DOCUMENT RESUME

ED 173 436 TM 009 575

TITLE	Proceedings of the Invitational Conference on Testing Problems (New York, New York, November 2, 1957).
INSTITUTION	Educational Testing Service, Princeton, N.J.
PUB DATE	2 Nov 57
NOTE	124p.
EDRS PRICE	MF01/PC05 Plus Postage.
DESCRIPTORS	*Achievement Tests; Cognitive Tests; Educational Testing; Elementary Secondary Education; *Evaluation Criteria; *Evaluation Needs; Higher Education; Humanities; Interest Tests; Natural Sciences; Personality Assessment; *Psychological Testing; School Personnel; Social Sciences; Student Testing; Talented Students; *Test Construction; *Testing Problems; Test Items; Test Results

ABSTRACT

This conference focused on the broad theme, improving the quality and scope of measurement. The first session centered on improving criteria for educational and psychological measurement, with papers on Criteria for Complex Mental Processes by Robert C. Wilson and on Criteria of Nonintellectual Aspects of Personality by Morris I. Stein. The second session considered the improvement of measurement through letter exercise writing, with papers on exercise writing in the field of the humanities (by Paul B. Diederich), in the natural sciences (by Leo Nedelsky), and in the social sciences (by Max D. Engelhart). The luncheon address on Prediction of Educational and Vocational Success Through Interest Measurement was given by Edward K. Strong, Jr. The final session concerned the application of test users' problems as guides to better measurement. Discussions included The School Administrator's Problems for Testers, by Paul T. Rankin; Discussion of the School Administrator's Problems, by Roger T. Lennon; The Guidance Director's Problems Suggestions for the Test Specialist, by Edward Landy; and The Consumer and the Producer, by Donald E. Super. (BH)

****
* Reproductions supplied by EDRS are the best that can be made *
* from the original document. *
****

ED173436

U S DEPARTMENT OF HEALTH
EDUCATION & WELFARE
NATIONAL INSTITUTE OF
EDUCATION

THIS DOCUMENT HAS BEEN REPRODUCED EXACTLY AS RECEIVED FROM THE PERSON OR ORGANIZATION ORIGINATING IT. POINTS OF VIEW OR OPINIONS STATED DO NOT NECESSARILY REPRESENT OFFICIAL NATIONAL INSTITUTE OF EDUCATION POSITION OR POLICY

1957

"PERMISSION TO REPRODUCE THIS MATERIAL HAS BEEN GRANTED BY
Educational
Testing Service

TO THE EDUCATIONAL RESOURCES INFORMATION CENTER (ERIC)."

TING

OBLEMS

TM009 575

EDUCATIONAL TESTING SERVICE

[illegible]

[illegible]

BOARD OF TRUSTEES, 1957-58

Katharine E. McBride, *Chairman*

Arthur S. Adams
Frank D. Ashburn
Grace V. Bird
Frank H. Bowles
Paul H. Buck
John T. Caldwell
James B. Conant
John W. Gardner
A. I. Henderson
Henry H. Hill
Frederick L. Hovde
Wallace Macgregor
Benjamin C. Willis
O. Meredith Wilson

OFFICERS

Henry Chauncey, *President*
William W. Turnbull, *Executive Vice President*
Henry S. Dyer, *Vice President*
Robert L. Ebel, *Vice President*
Jack K. Rimalover, *Secretary*
Catherine G. Sharp, *Assistant Secretary*
G. Dykeman Sterling, *Treasurer*
Robert F. Kolkebeck, *Assistant Treasurer*

Copyright, 1958, Educational Testing Service
20 Nassau Street, Princeton, N. J.
Printed in the United States of America

Library of Congress Catalog Number: 47-11220

INVITATIONAL CONFERENCE ON TESTING PROBLEMS

NOVEMBER 2, 1957

Improving the Quality and Scope of Measurement

ARTHUR E. TRAXLER, *Chairman*

Improving Criteria for Educational and Psychological Measurement

Improving Measurement through Better Exercise Writing

Test Users' Problems as Guides to Better Measurement

EDUCATIONAL TESTING SERVICE
PRINCETON, NEW JERSEY LOS ANGELES, CALIFORNIA

FOREWORD

Although the persons attending the Invitational Conference on Testing Problems have a common bond, they represent many different interests in the field of measurement. These range from the guidance counselor to the test developer, from the pure researcher to the school administrator, from the psychologist to the statistician.

This diversity of interest presents a real challenge to the Conference Chairman—how to structure a program that will at once hold the interest of this very heterogeneous group and yet enable each participant to carry away with him ideas, suggestions, and practices that are stimulating and useful?

The success of the 1957 Invitational Conference attests to the manner in which Chairman Arthur Traxler met this challenge. Beset with all the administrative problems of his own annual Educational Records Bureau Conference, he, nevertheless, succeeded in scheduling for the Invitational Conference, on the following day, a stimulating series of reports and discussions on the theme "Improving the Quality and Scope of Measurement." The large number of persons in the audience felt well-rewarded.

These few words cannot adequately express my appreciation for Arthur Traxler's efforts, for the excellent presentations of each of the speakers, and for the efficient manner in which the Conference was conducted. I hope these *Proceedings* will serve to give an even wider audience the full flavor of an important conference.

HENRY CHAUNCEY
President

PREFACE

The 1957 Invitational Conference on Testing Problems, which was held at the Hotel Roosevelt on November 2, 1957, was concerned with the broad theme "Improving the Quality and Scope of Measurement."

The improvement of criteria has long been recognized as one of the major measurement needs. In both the educational and vocational areas, demonstrated progress in measurement is likely to depend upon criteria of success more adequate than marks or over-all ratings. Where criteria are obscure in meaning or limited in scope, improvement in tests may not be reflected in higher correlations with the criteria. In fact, it is entirely possible for better tests to be related to a lesser degree with faulty criteria.

In an effort to stimulate thinking and discussion of this problem among test specialists, the first session of the Invitational Conference was concerned with the topic "Improving Criteria for Educational and Psychological Measurement," with papers on "Criteria for Complex Mental Processes" by Robert C. Wilson of Reed College and the Gifted Child Project, Portland (Oregon) Public Schools, and on "Criteria of Nonintellectual Aspects of Personality" by Morris I. Stein of the University of Chicago.

Another major need for progress in educational measurement is improvement in techniques of writing test exercises. This improvement may come about in two ways. In the first place, test constructors may devise new and novel situations on which test items are based. There might be more experimentation with listening comprehension, with laboratory demonstration followed by items calling for integration of ideas gleaned from observation, and so forth. In the second place, the quality of the conventional kind of item may be raised through improvement in the content and clarity of each item written and through ingenuity in devising some wholly different kinds of test items.

The second conference session considered the question of "Improving Measurement Through Better Exercise Writing," with stimulating papers on "Exercise Writing in the Field of the Humanities" by Paul B. Diederich of Educational Testing Service, "Exercise Writing in the Natural Sciences" by Leo Nedelsky of the University of Chicago, and "Exercise Writing in the Social Sciences" by Max D. Engelhart of the Chicago City Junior College.

A third measurement need, one which has been referred to with increasing frequency in recent years, is better communication between test specialists and test consumers. Test specialists need to convey the basic concepts of test construction, standardization, validation, and application in language that test users can understand. But, equally important, it is desirable to remember that ideas flow in both directions and that the measurement fraternity can learn much from the practitioners—the school administrators, the teachers, and the guidance counselors.

The afternoon session of the conference considered the topic "Test Users' Problems as Guides to Better Measurement." "The School Administrator's Problems for Testers" were presented by Paul T. Rankin of the Detroit Public Schools, and his paper was discussed by Roger T. Lennon of the World Book Company. "The Guidance Director's Problems and Suggestions for the Test Specialist" were presented by Edward Landy of the Newton (Massachusetts) Public Schools, with discussion by Donald E. Super of Teachers College, Columbia University.

An outstanding feature of the conference was the luncheon address by Emeritus Professor Edward K. Strong, Jr., of Stanford University, on "Prediction of Educational and Vocational Success Through Interest Measurement," a field to which Dr. Strong has devoted a lifetime of professional leadership.

The conference was attended by over 590 delegates, including a large proportion of the measurement specialists of the United States and a number from other countries as well.

ARTHUR E. TRAXLER
Chairman

CONTENTS

FOREWORD by Henry Chauncey 3

PREFACE by Arthur E. Traxler 5

SESSION I

"Improving Criteria for Educational and Psychological Measurement"

Remarks of the Chairman, JOSHUA A. FISHMAN, College Entrance Examination Board 11

IMPROVING CRITERIA FOR COMPLEX MENTAL PROCESSES
ROBERT C. WILSON, Reed College and Portland, Oregon, Public Schools 13

CRITERIA OF NONINTELLECTUAL ASPECTS OF PERSONALITY
MORRIS I. STEIN, The University of Chicago 21

SESSION II

"Improving Measurement Through Better Exercise Writing"

Remarks of the Chairman, J. WAYNE WRIGHTSTONE, New York City Board of Education 35

EXERCISE WRITING IN THE FIELD OF THE HUMANITIES
PAUL B. DIEDERICH, Educational Testing Service 36

EXERCISE WRITING IN THE NATURAL SCIENCES
LEO NEDELSKY, The University of Chicago 45

EXERCISE WRITING IN THE SOCIAL SCIENCES
MAX D. ENGELHART, The Chicago City Junior College 57

LUNCHEON ADDRESS

Chairman: HENRY CHAUNCEY, Educational Testing Service

PREDICTION OF EDUCATIONAL AND VOCATIONAL SUCCESS THROUGH INTEREST MEASUREMENT
EDWARD K. STRONG, JR., Stanford University 72

SESSION III

"Test Users' Problems as Guides to Better Measurement"

Remarks of the Chairman, ARTHUR E. TRAXLER, Educational Records Bureau 85

THE SCHOOL ADMINISTRATOR'S PROBLEMS FOR TESTERS
PAUL T. RANKIN, Detroit Public Schools 86

DISCUSSION OF THE SCHOOL ADMINISTRATOR'S PROBLEMS
ROGER T. LENNON, World Book Company 94

THE GUIDANCE DIRECTOR'S PROBLEMS AND SUGGESTIONS FOR THE TEST SPECIALIST
EDWARD LANDY, Newton, Massachusetts, Public Schools and Harvard University 100

THE CONSUMER AND THE PRODUCER
DONALD E. SUPER, Teachers College, Columbia University 112

APPENDIX 117

SESSION I

Improving Criteria for Educational and Psychological Measurement

Remarks of the Chairman

JOSHUA A. FISHMAN

Together with many of you I attended last year's Invitational Conference on Testing Problems. As always, this was a most rewarding experience and one which provided much food for further thought, discussion, and study after the Conference was over. The over-all theme of the Conference, as many of you will remember, was "Testing—Then and Now" and the program presented a magnificent panorama of the progress (and also, quite frankly, of the lack of progress) of the testing movement during the past several decades.

One of the major points implied by last year's Conference was the still far from satisfactory state of affairs—both conceptually and operationally—in conjunction with the entire area of test validity. Our major satisfactions in testing work must come from other sources.

During the past quarter century we have vastly improved our administering, scoring and reporting techniques. As a result, we can now calmly undertake testing programs of a magnitude that would have staggered the imagination only a few generations ago. Into the College Board testing program (with which I am most intimately familiar) a new test was introduced a short two years ago. In its very first administration it reached 100,000 individuals, and this year it has reached well over 300,000 individuals. (Both of these figures refer to a single day, nationwide administration calling upon the integrated cooperation of a veritable army of specialists of various kinds.) Certainly our ability to perform miracles of this kind has improved markedly in a relatively short time span. Our host at today's Conference is a prime example of truly unmatched and quite unbelievable accomplishment in this area—as in every area relating to testing.

A second major front of our progress has been in the crucial area of test reliability. Here we have perfected various conceptual models as well as operational procedures for appropriately estimating and successively improving the dependability of our measures. Our extensive work in scaling and in item analysis has slowly whittled away at test irreliability, especially in the classical achievement and aptitude areas, to the extent that test reliability is currently one of our chief sources of professional pride.

This brings us to the central issue of validity. Here our progress has been least satisfactory. The number of tests grows from year to year, the number of individuals tested rises continually, the reliability of the scores obtained is appreciable, and yet we are often embarrassed to admit that the criteria against which we validate these scores are frequently gross or ambiguous—at best partial, and at worst, non-existent. The correction of this state of affairs is a difficult task and one to which all of us—ETS, and everyone of us who builds or uses tests—must constantly readdress ourselves. This is, perhaps, the major testing frontier for the remainder of this century: the selection and combination of criteria appropriate to the subtle nuances represented by our test titles on the one hand and appropriate also to the complexities of the situations for which test behavior is supposedly predictive, on the other.

We are fortunate indeed to have an opportunity to listen to two speakers who have done considerable work in a research area for which the criterion problem is particularly difficult. Our first speaker, Dr. Robert Wilson, appears before us in the dual capacity of Associate Professor in the Department of Psychology at Reed College and Research Director of the Gifted Child Project of the Portland (Oregon) Public Schools. Dr. Wilson's topic is "Improving Criteria for Complex Mental Processes." Our second speaker, Dr. Morris I. Stein, is Associate Professor of Psychology at the University of Chicago. Dr. Stein will talk on "Criteria of Nonintellectual Aspects of Personality."

Improving Criteria for Complex Mental Processes

ROBERT C. WILSON

The inclusion of the topic "Improving Criteria for Complex Mental Processes" in today's program is in part an expression of the feelings of frustration experienced by test-developers. At least two factors have contributed to these feelings of frustration: first, the feeling on the part of test-constructors that they are actually improving the predictive power of their measures of aptitude and achievement, accompanied by an awareness that they have very little actual evidence that the improvement is substantial and a feeling that much of the difficulty is due to the crudeness and lack of sensitivity in such time-worn criteria as school or college grades; secondly, the criticisms expressed by test-users that the tests which are available measure or predict only *simple* mental processes and fail to supply sound measures of the more important kinds of mental functioning such as are involved in critical analysis, synthetic thinking and the various creative processes. These latter criticisms are in turn, perhaps, an expression of the feelings of frustration of test-users who are faced with the problems of measuring and predicting in complex situations.

May I add my own feelings of frustration as a test-constructor and test-user concerned with the problem of identifying "complex mental processes" in action and with the problem of observing them in a systematic manner. We have recently completed in Portland a five-year experimental study for the purpose of developing better educational experiences for both intellectually gifted and talented children. The results of these five years of experimentation have led to the development of a variety of special educational provisions for gifted children. These provisions were adopted last spring by the Portland School Board as a permanent part of the educational program of the Portland Public Schools and are being expanded to include all of the schools in the Portland Public School System (1).

The initial proposal to the Fund for the Advancement of Education in April 1952 requesting support for the study outlined five essential features:

1. Provision for many kinds of unusual ability so that the traits and talents selected for identification and for development shall not be limited to general intelligence as currently tested and shall

13

include creative, intellectual, artistic, and social capacities, and the emotional and moral qualities necessary for effective use of these capacities.

2. Experimentation with methods and materials of instruction for groups and individuals that will challenge and develop unusual abilities of various kinds, and to this end the encouragement and training of good teachers.
3. Coordination of the teaching and the programs of promising students with the common curriculum of the schools and with other educational resources in the community to avoid fixed grouping, with the intention of enabling other students, (and in some measure all students), to profit from the experimentation.
4. Cooperation with other colleges for following up the students from the program and for working out closer articulation of college curricula with those of the high schools, and with possible acceleration at either the high school or college level or both.
5. Close collaboration with a college of liberal arts and sciences in a strategic position for assisting in shaping and evaluating the program and for actively participating in important aspects of it.

The last three features of the program have been relatively easy of accomplishment and evaluation. The first two have been more difficult and relate closely to our topic today. Let us analyze them one at a time and expand their meaning. Proposed feature number one was: "Provision for many kinds of unusual ability so that the traits and talents selected for identification and for development shall not be limited to general intelligence as currently tested and shall include creative, intellectual, artistic, and social capacities, and the emotional and moral qualities necessary for effective use of these capacities." Let us omit the last two characteristics, "the emotional and moral qualities necessary for effective use of these capacities," since they fall in the realm of Dr. Stein's paper on, "Criteria of Nonintellective Aspects of Personality," and turn our attention to the first mentioned characteristics, that is, "creative, intellectual, artistic, and social capacities."

The identification of those pupils possessing the general intellectual qualities which are related to school success presented no great problem. We have used a number of standardized tests of aptitude and achievement in combination with teacher observations of classroom performance. We did encounter difficulties when we turned our attention to such complex processes as "creative, artistic, and social capacities."

It was decided at the beginning of the study that we would attempt to identify the ten per cent most talented students in each of seven talents. These were art, music, creative writing, creative dramatics, creative dance, social leadership, and mechanical talent.

To identify the special talents we established committees of teachers, administrators, supervisors, and lay people from the community who had technical training, ability, and interest in these talent areas. These committees examined the existing testing devices in the various fields, and in most cases found that there were no adequate testing devices available. The committees then proceeded to develop their own exercises for screening purposes in these various talents. In most instances they turned in the direction of developing performance tests. We now have available seven sets of screening exercises, one for each of the talent areas mentioned earlier. These exercises are not all given in any single grade, rather they are distributed over several grades to avoid overloading any single grade. However, the exercises for each talent are administered in at least two grades. By way of example, I will describe one of the screening devices, that for creative writing, in some detail.

The creative writing screening device consists of five exercises which are administered to all fifth and sixth grade children. The first exercise is concerned with "Developing Expressive Sentences." The teacher gives the children some preparation for this exercise and then on the day of the test gives the children sentences like this, for example: "The man went down the street." The children are asked, "In what way could you add to or change the word man to give a clearer picture of the man? In what ways could you change other words in the sentence to make us see this man going down the street?" Here is an example of what one child did with this, he changed "the man went down the street" to "the blind man with the white cane tapped his way down the busy street." The children are given several such sentences to develop and expand in order to make them more expressive. The results are rated by the teacher on a five point scale in terms of certain criteria.

The second exercise is "Developing a Paragraph from a Sentence." The children are given some preparation for this and then are given several sentences and told to choose one and write a paragraph about it. For example, "The mysterious box drew all eyes to it," or "After all it was only a mouse."

The third exercise is "Writing a Story from Descriptive Phrases." The teacher writes on the board several descriptive phrases, such as, "high-fenced yard," "yelling crowd," "abandoned mine," "stormy sea," "faithful horse," etc. The children are to select three of these phrases and develop a story using the descriptive phrases in the story.

In the fourth exercise, "Writing an Experience," the children are asked to tell the most exciting or amusing experience, imaginative or real, they have had in the past few months.

The fifth exercise is writing an imaginary conversation between two characters from fiction the child has read.

Exercises similar to these are used in several of the other talent areas. They are generally rated by the teacher, which has some limitations but also some advantages. I would like to quote from a statement of the philosophy behind these screening exercises: "Since the number of talented children will vary from class to class, and since the standards of rating will vary from teacher to teacher, it is impossible to specify that a particular number of children be regarded as talented or that a particular score be regarded as the minimum cutoff. As a general guide, however, it is likely that on the average there will be one to three outstanding pupils in a class. In some classes, however, there will be none and in some exceptional classes there may be as many as six or eight. In looking at the total scores, the teacher may notice a few high scores close together with a larger gap between these and the next highest scores. When this is the case, the gap may be used as the cutoff point and the children above this point may be tentatively identified.

"In interpreting the results of the talent screening devices it is important to bear in mind their purpose and the uses to which they may be put. Their purpose is to help us pick out those children who may benefit most from further work in the talent being tested. They may be used to enable us to achieve more closely the ideal which is often expressed of 'meeting the needs of each child.'

"The purpose of the identification exercises is *not* the assigning of labels. They should by no means be regarded as a certification device or a stamp of approval. The screening devices and the administrative procedures used are only a rough screening procedure. They are far from perfect. We may identify as promising some children who will not live up to this promise in further work. We may also miss a few children who actually do have promise.

"Our hope is that by providing a standard set of exercises, all children will have an opportunity to show their abilities and that those selected as outstanding may have an opportunity for further development of their abilities. By repeating the exercise at several grades, we may have increasing confidence in our selection if a child continues to show up well."

The preceding is an example of one attempt to develop tests for certain complex mental processes—the development of informal tests disguised as ordinary classroom exercises which may be used to elicit teacher judgments about who should receive special attention in each of several areas. This has proved to be a useful approach. Pupils selected for special classes on the basis of these screening exercises do well in the classes.

Our second major objective included the statement, "Experimentation with methods and materials of instruction for groups and individuals that will challenge and develop unusual abilities of various kinds." A variety of special provisions have been tried out, and I do not have time here to describe them all.

One of the methods taken to elaborate this objective was the establishment of seminars in the high schools. Seminars are offered in literature, social studies, mathematics and science. The seminar is here defined as a small group of students, with one or more teachers, following an unprescribed, elective course, the content of which goes above and beyond the offerings of normal classes and emphasizes self-directed study, the critical examination and interpretation of common readings, and penetrating class discussion. Among the aims of such seminars are those of: (1) making students more intelligent and perceptive readers, (2) improving student writing, (3) developing the ability to do independent research work, (4) developing the ability to think critically, and (5) developing the ability to think creatively. These all involve complex behavior and complex mental processes. While we have used a variety of formal and informal means of evaluating the achievement of these aims, we are not satisfied with the techniques we have used.

Turning now from our own particular problems, in Portland, I should like to indicate two major directions in which I believe progress will be made in improving our understanding of and our ability to measure complex mental processes and some examples of work now being done in these directions.

One of these directions is that of *clarifying our conceptualizations of the complex mental processes* so that the tests we develop of these processes will have greater content validity and will elucidate our understanding of human mental behavior. The other direction is that of *specifying more carefully* the *situational factors* which relate to the predictive validity of any particular application of a testing device—an elaboration or analysis of the multi-determined criteria which are what we usually have available.

In regard to the first direction, one may observe that much progress in science has resulted from the clarification of concepts through the substitution of more precise terms for certain commonsensical, popular or global terms. Two efforts which seem to me to offer useful possibilities in the development of new concepts and new distinctions are those reported by Guilford and Bloom.

Guilford has been directing since 1949 a series of studies of aptitudes of high-level personnel under contract with the Office of Naval Research. These studies have had as their starting points four global and commonsense concepts—reasoning, creativity, planning, and evaluation. The general plan of attack has been, first, to develop hypotheses concerning more specific abilities which might go into these global concepts. Then a variety of tests are developed to measure alternative conceptions of the various specific abilities. The tests are then administered to large numbers of individuals and the results factor analyzed, leading to the acceptance or rejection of alternative hypotheses. During the eight years in which these studies have been going on a number of factors discovered by other investigators have been confirmed and a number of new ones have been discovered. In all, about forty factors seem well established.

Guilford has noted certain regularities in the tests of these factors and on the basis of them has developed a theoretical model of intellect. This is described in detail in two of his recent publications (2, 3). In brief, he first divides intellect into *memory* and *thinking*. He further divides the thinking abilities into three categories under the headings of *cognition*, *production*, and *evaluation*. The cognitive abilities are defined as those which have to do with the discovery of information; the productive abilities have to do with the use of information; and the evaluative abilities have to do with decisions as to the goodness, accuracy, or suitability of information or products. He further subdivides the abilities listed under *production* into those involved in thinking that converges on one right answer and those involved in thinking that diverges or goes off in different directions. The latter is most closely related to creative thinking.

The tests under all of these classifications are further categorized in terms of their figural, structural, or conceptual properties. Guilford's studies have already provided a wide variety of new tests many of which are presently being used by other investigators. The theoretical model is suggestive of still other tests and abilities and further studies which may clarify our understanding of complex mental processes.

A second effort, which I would like to mention, in the development of new concepts and new distinctions among complex mental processes is the work of a committee of college and university examiners. The results of their deliberations have appeared in a book entitled "A Taxonomy of Education Objectives, Handbook I: Cognitive Domain" edited by Benjamin S. Bloom (4). One of the purposes of the book is to provide a hierarchical classification of measurable educational objectives in the cognitive area. Cognitive is used to include activities such as remembering and recalling knowledge, thinking, problem solving, and creating.

The book includes descriptions in terms of changes in student behavior of six major categories of educational objectives in the cognitive domain. The lowest category is *knowledge* defined as the recall of specifics and universals, the recall of methods and processes, or the recall of a pattern, structure, or setting. Parenthetically, it is recognized that this type of educational objective is the easiest to measure, and therefore the one most often emphasized in current curricula. The category at the next lowest level is called *comprehension* which is defined as including those objectives, behaviors, or responses which represent an understanding of the literal message contained in a communication. The next level is called *application* and involves the use of abstractions in particular and concrete situations. At the fourth level is a category called *analysis* which involves the breakdown of a communication into its constituent elements or parts such that the relative hierarchy of ideas is made clear and/or the relations between the ideas expressed are made explicit. The fifth level is called *synthesis* which is defined as the putting together of elements and parts so as to form a whole, and the sixth level is called *evaluation* which involves judgments about the value of material and methods for given purposes. Each of the major categories is divided into sub-categories on the basis of observable differences in student behavior.

The book includes descriptions of sample objectives and activities under each category, also sample test items for attempting to measure each category and a discussion of the characteristics of test items appropriate to the measurement of each category of objectives.

Among the outstanding features of this conceptualization of complex mental processes are, first, its emphasis on changes in pupil behavior. It is hoped that this will exercise a desirable influence on educators in terms of encouraging them to formulate their objectives in more behavioral and measurable terms. Second, it is suggestive of new research in analyzing and measuring some of the suggested categories. Third, it should facilitate communication among educators and testers.

I shall mention only briefly the second direction in which I believe progress will be made in improving our understanding of the complex mental processes, that is the direction of specifying more carefully the elements in the criterion which may affect the predictive validity of any particular application of a testing device. Usually the criterion measures which we have available are complex in that they reflect many elements.

One example of efforts to determine the components of complex criteria is the application of factor-analysis to correlation matrices which include both criterion measures and a variety of factored tests. The loadings of a criterion measure on the obtained factors yield useful information concerning the nature of the criterion.

A second example of efforts to determine the components of complex criteria is found in the work of Stern, Stein and Bloom as elaborated in their book "Methods in Personality Assessment" (5). The methods described have implications for both the intellective and non-intellective aspects of personality. I hope that our next speaker will elaborate on these methods further.

REFERENCES

1. *A report summarizing four years of progress by the Cooperative Program for Students of Exceptional Endowment.* Portland, Oregon, Public Schools (mimeo), 1956.
2. GUILFORD, J. P. The structure of intellect. *Psychol. Bull.*, 1956, *53*, 267-293.
3. GUILFORD, J. P. A revised structure of intellect. *Rep. psychol. Lab.*, No. 19. Los Angeles: Univer. of Southern Calif., 1957.
4. BLOOM, B. S., et al. *A taxonomy of educational objectives.* New York: Longmans Green, 1954.
5. STERN, G. G., STEIN, M. I., AND BLOOM, B. S. *Methods in personality assessment.* Glencoe, Ill., Free Press, 1956.

Criteria of Nonintellectual Aspects of Personality

MORRIS I. STEIN

When I first saw the title of the talk that was assigned to me—"Criteria of Nonintellectual Aspects of Personality," I was fascinated by the prefix "non" in the word "nonintellectual." It reminded me of an old folk-tale of how the first census was taken. Many long years ago, the story has it, people became interested in learning how many of them there were. They lined themselves up, appointed a census taker, and he counted as follows: "not-one," "not-two," "not-three," and so on. This curious counting procedure obviously had a purpose. The people of the period were quite superstitious and they feared that if they counted in a straightforward fashion—"one," "two," "three," then the devil would know how many people there were and he would be aware of how many souls he still had to destroy. Nevertheless, the people wanted to know how many of them there were and, to fool the devil, they counted "not-one," "not-two," "not-three."

As you might have inferred from the folk-tale, I believe that personality factors have not received the attention they deserve in the field of measurement. Indeed, in some quarters, they have been relegated to second-class citizenship because they *are* nonintellectual factors. Some people shy away from this field because they believe that it is all confused. And when I talk to these people I often wonder what their positions would have been if they, rather than Binet, had been asked to measure intelligence. How would they have defined intelligence? How would they proceed with this problem which is really no more nor less complex than the measurement of personality factors. There are other people who have become so impressed with the failures reported in certain large-scale assessment studies that not only are they ready to throw out assessment but also the techniques that have been used, even though the techniques were not really put to test in these studies.

The field of measurement can no longer afford to hide behind rationalizations or behind professional bias. Measures of intellectual functions have not given us all that we desire in terms of understanding and predicting behavior. If you wish, the devil will have his due. It is no longer a question of whether or not one likes or wants to study personality factors. They must be studied. It is no longer a question of whether the problems in this area are difficult. They are.

In the time allotted me I shall assume that we are well agreed as to the importance of studying personality factors and I shall turn to a consideration of some major problems that are critical obstacles in the way of progress.

A Criterion

What is a criterion? There are two major ways in which this term has been used and it is necessary to differentiate between them for purposes of our discussion.

1. A criterion is a term assigned to a class of an individual's behaviors or responses. This I shall refer to later as "class criterion."
2. A criterion is the term assigned to a standard of performance. This I shall refer to later as "performance criterion."

In the first, "class criterion," the primary purpose is to separate classes or groups of individuals and in the second, "performance criterion," the primary purpose is to study the relationship between the basis for separating groups of individuals and some judgment or evaluation. The first is related to construct validity and the second is related to concurrent and predictive validity. In the first, a typical question is, "How can one separate persons in terms of anxiety, ego strength, need achievement, need dominance, etc.?" In the second, a typical question is, "What is the relationship between anxiety and good performance in a problem-solving situation, good performance as a clinical psychologist, good performance as an executive, and good performance as a company employee, etc.?" Since some of the problems involved in class criteria are different from those involved in performance criteria, I shall discuss each of them separately.

Class Criteria

In looking over personality tests in terms of class criteria, there are two types of problems. One type of problem is inherent in the difficulty of the task and the other type of problem seems to be generated by psychologists themselves and is in the nature of a communication difficulty. Both these problems overlap a great deal, as you will observe in the discussion that follows.

One of the biggest problems in the measurement field is that psychologists have not yet agreed upon a taxonomy of personality characteristics. (Hopefully, we might follow Bloom's work as Dr. Wilson has suggested.) There are some who concern themselves with phenotypic characteristics while others concern themselves with genotypic characteristics. One investigator considers "leadership" a personality characteristic while another considers "leadership" the outcome between an individual's needs, etc. and the specific situation in which he finds himself. This is one type of confusion. Another type of confusion arises when several investigators use the same term for a personality factor but define it differently. Both McClelland and Edwards have tests in which need achievement is a variable but both investigators define need achievement in different ways. Low correlations between the two tests are not only a function of the different techniques that are used to get at need achievement but also because different types of need achievement are studied as suggested by Bendig (1) who found a correlation of .11 between Edwards' and McClelland's tests. These are two types of confusions that psychologists themselves are responsible for. The difficulties do not lie in the personality factors studied but in the people studying them. To exaggerate the case a bit, psychologists have created their own Tower of Babel. At times it appears as if what is necessary is the creation of a technical semantic society under the auspices of the American Psychological Association or some other group that would be charged with the responsibility for providing us with a standard definition of terms. As each new investigator came along and selected a variable for study he would accept the standard definition or else show good cause why it should be altered.

Developing and accepting a taxonomy of personality characteristics would have an additional advantage. It would help make us aware of the gaps in our knowledge. It would alert us to those personality factors that require more study and keep us from being redundant in our research.

A second type of problem resides more in the nature of the task but here too psychologists have added to the confusion. Assuming that an investigator has selected a personality factor for study, he is then confronted with a rather large universe of behaviors and responses from which he may wish to select a sample that would be appropriate to the selected construct. In the area of anxiety, for example, the investigator is confronted with the following possibilities: physical symptoms, functioning under experimental stress, behavior in social situations, defense mechanisms and the like. For ego strength, Murray and Kluckhohn have listed the following three major areas: perception and apperception, intellection and conation and their fifteen sub-areas—external objectivity, internal objectivity, long apperceptive span, concentration (directionality), conjunctivity of thought and speech, referentiality of thought and speech, will power, conjunctivity of action, resolution of conflict, selection of impulses, selection of social pressures and influences, initiative and self-sufficiency, responsibility for collective action, adherence to resolutions and agreements, and absence of pathological symptoms (2). Confronted with all these possibilities, no single investigator can, at any one time, study each major concept as thoroughly as it should be studied. Each investigator obviously selects that behavioral constellation which interests him and then investigates it as thoroughly as possible. To be sure, problems are encountered. On a formal level, however, they are no more severe than the problems involved in achievement testing. In achievement testing the examiner carves out a specific area for his study, e.g., 4th grade arithmetic. He does not investigate the child's ability in mathematics. In the personality area, however, there are relatively few instances where a narrow field is studied. Personality people tend to strive for the broader concepts and the global terms. Here is where psychologists add to the confusion. After they have developed a test or a technique for studying a specific aspect of behavior, they entitle their tests with global terms and thereby add much surplus meaning. Once again consider the problem of anxiety. There must be any number of tests that purport to measure anxiety but what kind of anxiety are they measuring? Are they measuring free floating anxiety, bound anxiety, guilt anxiety, shame anxiety, or what? A recent factor analysis of the Taylor Scale (3) indicates that it measures at least five different kinds of anxiety.

This is the natural course of development with most any test, whether it concerns itself with cognitive or personality characteristics. After the originator of the test has presented his results, others help by clarifying its purposes and areas of greatest efficiency. But what also seems to happen is that a test gets fixed with its original title and the unsuspecting investigator who wants a test of anxiety picks up the one that is most readily available or the one that is most featured in the recent journals and uses it without investigating the type of anxiety the test supposedly measures and whether that type of anxiety is relevant to his purposes. Indeed, this is a marketing problem and we can always say "*Caveat emptor*," but I think we have to assume more responsibility than that. The Technical Recommendations of the A. P. A. committee may help immeasurably in this regard. The only thing I would have to add to these recommendations in the light of what I have just said is that each test in the personality area be entitled by both its class and sub-class terms. Thus, in the area of anxiety, no test would be published simply under the heading "Anxiety" but under the heading "Anxiety—Physical Symptoms"; "Anxiety—Guilt"; "Anxiety—Shame." In this fashion not only would we clarify what we are talking about but we would also curtail some of the confusion that arises when cross-validation is attempted.

A third type of problem in the field of personality measurement is the question of what type of technique should be used in presenting the test stimuli to the subject. At the present time a large variety of techniques do exist. There are paper and pencil tests, sentence completion tests, story tests, ink-blot tests, drawing tests, situational tests, experimental measures, physiological measures, etc. Indeed, psychologists have shown ingenuity in technique development. Since man is a complex and multi-faceted organism who functions on at least two levels—conscious and unconscious—such developments are necessary and no doubt valuable. But the problem is that we do not understand the relationship between these techniques and relatively little time and energy is devoted to understanding these relationships. Here, I would state the problem as follows: Since man is a complex organism how do the character of the data obtained from him with one method compare with the character of the data obtained from him with another method. In some instances there may be high positive relationships, zero relationships, or even negative relationships. One should not be either encouraged or discouraged by the magnitude of the relationship but rather be curious about why the relationship has occurred at all. Typically, however, if a negative relationship is found between two personality tests it is more likely than not that one of them gets thrown out and the reasons for the negative relationship are not at all investigated. To take an extreme analogy, it is as if the investigator were the first person ever to study the oxygen content of the blood and, after drawing two blood samples—one from the veins and the other from the arteries—he throws up his hands in despair because he has found a negative correlation.

There are at least two reasons why the problem I have posed has not received the attention it deserves, and both of them have nothing to do with the criterion problem but with psychologists. First, the professionalism in our field has resulted in the development of specialists. There are Rorschach'ers, TAT'ers, paper and pencil testers and the like. These specialists even have their own societies and they have achieved a level of trained incapacity so that they find it difficult to communicate with their colleagues. Consequently, the focus of attention is no longer on the problems in the field of personality, but on techniques of obtaining personality test data and scores. The second reason why this problem has not received more attention is that psychologists have tackled the prediction problem even before they coped with the understanding problem and, when the predictions failed, they were all too ready to discard the tests. The Rorschach and other projective tests have been used in a number of predictions and assessment studies and why they were used with the expectation that they would be excellent predictors when we really did not have a complete grasp of the factors involved in them, is rather difficult to fathom. After the first few experiences, there should have been sufficient evidence as to the difficulties involved but the experiences were permitted to accumulate. I would suggest that we bring the problems back into focus and the techniques will take care of themselves or they will be developed as the occasion demands.

A fourth problem in this area is the population selected for study. Let us assume that a researcher desires to investigate the relationship between personality factor "x" and something else called "y". He would obviously do well to have some *a priori* basis for assuming that factor "x" exists in his population and that it exists with sufficient range to make the study worthwhile. This seems like an obvious procedure but there are times when the investigator studies an available population, usually the ubiquitous college sophomore, on the assumption that personality factors are normally distributed. Then, when he investigates the relationships between factor "x" and "y" he either finds negative results or he does not replicate the work of others. Much time and effort could have been saved if an *a priori* analysis or even a pilot study had been undertaken. Lest you misunderstand, the difficulty here is not one of experimental design or the value of a pilot study but, rather, a way of thinking.

The limitations of time do not allow me to do more than treat in a rather cursory fashion one of the issues involved. In much personality research we concern ourselves with the relationship between a single personality variable and some other variables. The fact of the matter is that the human organism is not made up of independent variables but of interdependent variables. For example, to take a single case, the individual who scores high on a test of "anxiety—free floating" and high on a test of "ego-strength—conjunctivity" is different from one who scores high on the test of anxiety but low in ego-strength. If we limit ourselves simply to the anxiety, our experimental results may be quite limited. Restating this issue in other terms, it may be said that one of the difficulties in working with non-intellectual factors as well as with the intellectual ones is that we have not dealt adequately with the "organismic" and typology problems. Indeed, these are difficult problems but these difficulties should not deter us from seeking a solution. The fact that we have been dissatisfied with previous attempts at typological systems has deterred us as has the belief that all people are equal in terms of personality. Probably solution of the problem will be more feasible now that techniques such as latent structure analysis are available. But the technique itself can only complement some good theorizing.

The four problems I have just discussed are not the only problems in dealing with class criteria but my analysis of them leads me to the conclusion that there is no "criterion" problem but rather a problem in developing order out of disorder. This goal will no doubt be achieved if we declare a moratorium on the internecine conflicts between psychologists and become more problem-oriented.

Performance Criteria

In the time that remains, I should like to say a few words about performance criteria.

As psychologists turn to the prediction of behavior in complex social situations, they are often discouraged by the fact that some large-scale assessment studies have failed. It is often suggested that these studies failed because of difficulties involved in the "criterion problem." I would like to suggest that these studies failed not because of the "criterion problem" but because the researchers did not differentiate between a criterion and a standard of performance. When the assessors tried to pick the "good clinical psychologist," the "good researcher," or the "good psychotherapist," they overlooked the significance of the word "good" and the fact that *people* were making these judgments. In other words, the assessors were not studying "*the* good clinician"—they were studying men who were *regarded* as "good" or "poor" by a group of significant others. The judges, by their decisions, had signified who had and who had not attained a certain level of competence. These standards of performance are not psychological variables; they are social evaluations. The psychological variables or what I would call the criteria, are those which the assessors assume to be involved in fulfilling the demands of the situation. It is the assessor's task to "get behind" the judges' verbal statements and determine their bases for evaluation. Having done this, he may then be able to decide on the psychological variables he will investigate and the instruments or techniques he will use to obtain his data.

I can perhaps make my point a little more clear by citing an experience during the course of the research that resulted in the book *Methods in Personality Assessment*, by Stern, Bloom and myself (4).

We had been asked by the Dean of a school to select the creative student in his area. During the course of some preliminary discussions, the Dean was asked to describe the creative student. He started by saying, "He has a crew cut, tweed jacket, flannel trousers and saddle shoes." The Dean had ranked his ten students in terms of creativity and the assessment group studied these students with a variety of psychological tests. The assessors' rank order of the students matched perfectly with the rank order obtained from the Dean. The Dean was overjoyed that he now had a series of techniques on which to base his future judgment. To be sure, the techniques were available but he did not have what has been called the criterion. The criterion, in many respects, was a very simple one. After the Dean had described the creative student for us and since there was such a close match between his description of the student and his own appearance, it became apparent to the assessors that the creative student for this Dean would be one who was created in his own image and, therefore, one major issue for the assessors was to determine what kind of a person the Dean was and who could get along with him.

This example illustrates what I mean by getting behind the verbalization of the judges. The problem here is not a simple one, to be sure, but it can often be resolved if the psychologist spends time in carefully observing his subjects while at work or in extended discussions with the judges. The psychologist's task is also facilitated if he makes use of methods and techniques used in the other social sciences. For example, in my research on creativity, we have found it quite valuable to analyze the environments of our researchers in terms of roles and for two of these, the scientific and professional roles, we were able to draw quite heavily on the analyses already available from the sociologists.

In *Methods of Personality Assessment* we have also suggested the value of developing the models of personality necessary to fulfill the environmental demands and we have also indicated several approaches for arriving at and clarifying these models.

Comparing the discussion of class criteria with the discussion of performance criteria, it is apparent that the role of the psychologist in both situations may in certain respects be quite different. In general, when dealing with performance criteria, the classes of subjects are determined for him by the judges and more often than not they are not psychologists. When dealing with class criteria, the psychologist generally determines his own classes or works with someone who has been trained in this area. Some psychologists often find it difficult to accept the role demands in the assessment situation. That is, when the psychologist is asked to select the "good clinician," "good leader," etc., he often has ideas of his own as to what such a person should be and he is likely to select these variables for testing. But, unless his thoughts are congruent with those that exist in the minds of the person making the judgments, he is likely to fail. In the assessment situation he is best off if he is quite realistic and determines the requirements of the field before he selects his variables and techniques.

In concluding this section on the criterion as a standard of performance, I can only repeat Lewin's formula BPE—Behavior is a function of the transactional relationship between the person and his environment. Although we still have much to learn about the person, it is incumbent upon us also to understand the environment or else our predictions will fail.

In summary then I have presented a series of problems that confront us when we deal with criteria. These problems are essentially not very different whether we consider the intellectual or nonintellectual aspects of personality. While I may have raised more issues than I have answered, I cannot help but feel that they will be resolved as we devote more time to analyzing them and to clarifying the basic issues.

REFERENCES

1. Bendig, A. W. Manifest anxiety and projective and objective measures of need achievement. *J. Consult. Psycho.*, 1957, *21*, 354.
2. Murray, H. A., and Kluckhohn, C. Outline of a conception of personality. In Kluckhohn, C., Murray, H.A., and Schneider, D. M. (Eds.) *Personality in Nature, Society and Culture* (Rev. Ed.), New York: Knopf, 1953, 3-49.
3. O'Connor, J. P., Lorr, M., and Stafford, W. Some patterns of manifest anxiety. *J. Clin. Psychol.*, 1956, *12*, 160-163.
4. Stern, G. G., Stein, M. I., and Bloom, B. S. *Methods in Personality Assessment*, Glencoe, Illinois: Free Press, 1956.

DISCUSSION

Participants: Anne Anastasi, William E. Coffman, Joshua A. Fishman, Leo Nedelsky, Morris I. Stein, Robert C. Wilson.

DR. NEDELSKY: Dr. Stein believes that a psychologist's description of his own work as a study of anxiety is not precise enough; the description should indicate the type of anxiety studied, e.g., anxiety-shame or anxiety-guilt. I don't know whether psychology has reached the stage of development at which mental constructs such as anxiety-shame—and I take it that the concept is a mental construct and not a directly observable phenomenon—are so soundly based and firmly established that *all* psychologists must use them as terms of analysis. In physics, there was a time, a very productive time indeed, when many key terms and concepts were fluid and equivocal. The seventeenth and eighteenth century physicists in their search for concepts that are basic did not agree on the meaning of mass, heat, quantity of motion, force. Instead, they went their separate ways and used these terms tentatively, e.g., Huygens used mass and size interchangeably, and, as tools of analysis, in a way that seemed most fruitful to each investigator. It was not till physics was considerably more advanced that the meaning of these concepts was crystallized and set. In contrast, some centuries earlier, there *was* a set of well-agreed on mental constructs—the four elements: earth, water, air, and fire, of which all things were thought to be made. I can imagine a paper presented at the University of Paris in the Middle Ages being turned down for "lack of clarity and precision" because it described a new substance without specifying the proportion of the four elements of which it was composed.

CHAIRMAN FISHMAN: I wonder if either of the speakers would care to react to this?

DR. STEIN: Dr. Nedelsky has certainly pointed to an important problem. The scientist's language and his classification system may stand in the way of progress. Unless the scientist takes time out to question and test his classification system, he will limit himself only to those data for which he has a structure. In my talk I tried to make a plea for order, not constricting order.

DR. ANASTASI: I, too, should like to direct a question to Dr. Stein. It is my impression that one of the principal objectives of factor analysis is taxonomy. I don't mean to imply that all factor analysts have been successful in this regard, but I am wondering how Dr. Stein feels about that methodology, and how he feels about it in comparison with other methodologies in arriving at taxonomy.

DR. STEIN: I have high regard for factor analysis. I fear, however, that we may have permitted our statistical techniques to affect and direct our thinking beyond what is desirable. I do not object to statistical techniques but to psychologists who use them instead of thinking. Rather than sit down and theorize about personality or the personality factors involved in a research, some psychologists tend to jump to writing items (because it is easy to do so) and then use statistics to find out what they mean. This can result in empirical relationships that may be difficult if not impossible to explain. I guess that this is a statement of faith that I do believe that the problem of research should first be analyzed in terms of the personality factors involved, then the appropriate tests or items should be selected, and finally, the results subjected to statistical analysis.

DR. COFFMAN: I would like to ask Dr. Wilson a question.

You indicated that in identifying the talented, you were concerned with other characteristics than the intelligence as measured by commonly used tests. Have you made any study? Have you given any attention to what I presume is a group of youngsters in this situation who are not identified by your various methods of defining talent, but who made very high scores on intelligence tests?

DR. WILSON: Do we have students who score high on intelligence tests but do not score high on any of our talent devices? We have not singled out such students for particular study. Most of these talents that I have mentioned are in the various arts. We find that they are somewhat independent of standardized tests of intelligence in that there are many students, who are quite talented in these special talents, who may not be very bright.

CHAIRMAN FISHMAN: The question was referring to the reverse situation. Do you want to follow it up a little further?

DR. COFFMAN: Yes, I was just thinking whether ten years from now somebody might not be wondering whether they had missed a bet in not paying some attention to these bright, "untalented" youngsters.

DR. WILSON: I don't even know if there are such, or what percentage of the population they constitute—I expect a rather small one. I didn't mean to give the impression that we are only interested in the special talents. We do, *of course*, identify students on the basis of tests of intelligence and scholastic aptitude and the major part of our program is concerned with making better educational provisions for such students. I emphasized the creative areas in this talk since they are the areas which are less commonly dealt with in programs for the gifted.

CHAIRMAN FISHMAN: This is really a very serious problem from any point of view—the point of view of just humanity, or point of view of natural resources—individuals of very superior ability whose talents we can not discover. There is a rather serious area for research.

UNIDENTIFIED SPEAKER: Question for Dr. Stein.

You were referring to the disappointing results obtained in very particular studies using the tests against the usual criteria we have, and you suggested that the reason for this might be because we have too little understanding about the nature of man.

I would like to suggest for your consideration that one of the ways we have of achieving that is to get some measure of relationship among our criteria and reference variables. This is an approach which in the past has yielded some useful purpose. And I feel that while a test which shows disappointing results against some sort of criteria shouldn't necessarily be the reason for sorrow, neither should it be the reason for rejoicing.

DR. STEIN: The point I tried to make was that if we really clarified our criteria, we might have more occasions for rejoicing. To regard "good student" or "good clinician" as the criterion without understanding the psychological factors involved in arriving at the "good" evaluation will certainly limit us in our attempts to make valid predictions.

CHAIRMAN FISHMAN: Our time draws to a close. I think you will agree with me that this just scratched the surface of what might be involved in this area. Certainly, both of our speakers have emphasized more the conceptual angle (which might include the semantic and the various communication problems) in refining our thinking in this area, rather than any of the other possible approaches to criteria, more ultimate or more behavioral approaches to what we have in mind with our tests. But this is the kind of topic we are going to have to return to many times.

SESSION II

Improving Measurement Through Better Exercise Writing

Remarks of the Chairman

J. WAYNE WRIGHTSTONE

Although the words "exercise writing" in the title of the second session, "Improving Measurement Through Better Exercise Writing," are somewhat ambiguous, most of us here today, I am sure, realize that exercise writing is intended to mean the formulation of *objective* test exercises. We are thinking in much broader terms than in item writing because we are concerned with the total situation on which the items are based, as well as the items themselves.

It is our hope that the type of exercises that will be discussed at this session will be the kind designed to tap some of the more complex mental processes rather than recall or recognition of facts alone. Moreover, we hope that student answers, or responses, can be recorded so that scoring will be objective rather than essay examination type answers, which are frequently associated with the term "exercise writing," and require more subjective scoring procedures.

Exercise Writing in the Field of the Humanities

PAUL B. DIEDERICH

I am indebted to the next speaker, Leo Nedelsky, a former colleague in the Office of the University Examiner at the University of Chicago, for an illustration of a classic type of test-exercise in the humanities. When he graduated from high school as a refugee from Russia after the Revolution, there were only 40,000 such graduates in all Russia; hence the final examination was quite an occasion. Candidates had to appear in a public place several hours a day for a whole week and answer any questions that the townspeople wanted to ask. Leo was almost caught by the Archbishop of the Orthodox Church, a kindly old man then in his ninety-third year. After the examination had proceeded for some time, the Archbishop indicated that he would like to ask the young man a question. "Young man," he said, "when the soldiers opened Christ's side with the lance, why did both water and blood come forth?" Leo was stuck; he was not prepared for this question. But as he put his head down and appeared to be thinking furiously, it occurred to him that the old man was really in his dotage, and if he could only postpone his reply for about a minute, the Archbishop might forget what he had asked. He therefore continued thinking, and after a suitable pause, raised his head as though inspiration had struck and replied, "Your Reverence, when the soldiers opened Christ's side with the lance, *both* water and blood came forth!" "That's right," said the Archbishop benignly. "A prize pupil!" Leo said that, as he glanced around the table, he saw a black-bearded Jesuit gazing intently at him as though to say, "We need that boy in our Order!" It is something to be said for this form of examination, that, although Leo might have had theological objections, that Jesuit was undoubtedly right.

For all its merits, this form of test-exercise has all but disappeared except in the archaic ceremonial of the oral examination for the Ph.D. Coming down closer to our own time, the sort of test-exercise that is still most commonly used in the humanities is well illustrated by a story that my old professor of the Bible at Harvard, Kirsopp Lake, told in one of his last lectures. It was the day before our final examination, and I think he tried to ease the tension by saying, "Gentlemen, I had a remarkable dream last night. I dreamed that I was sitting on a cloud at Judgment Day, watching all the tribes of earth assemble. They all came together into a great plain and sat down. Then, out of the circumambient mist, a great hand arose and began writing on a celestial blackboard in letters large enough for all the world to read. It wrote out the Ten Commandments, and then—in typical examination fashion—it added: 'Students choose six.' "

These two types of questions pretty well cover the history of examining techniques in the humanities down to our own time. To carry on the story from that point, I obviously have an enormous field to cover in a very brief time. It usually takes me a semester. With your permission, therefore, I shall pass over the fields of art and music with only a brief comment on each, chiefly because they require so much testing time per item before I could make them come alive for you. I should have to have at least tape recordings and slides, and I prefer a live pianist and original works of art loaned from some museum. If I had the pianist, it would not be difficult, but it would take a good deal of time, for her to play snatches of twenty very familiar melodies, like "My country, 'tis of thee," and "Old Black Joe." In about half of these excerpts I should coach her to insert one definite but not too obvious error. You would mark each excerpt C if it was correct and E if it contained an error. If I wanted to increase the difficulty, I might ask you to indicate what was wrong with each excerpt that you marked E: to mark it 1 if the error was in the melody, 2 if it was in the harmony, 3 if it was in the rhythm, and 4 if it was in the expression or emphasis—that is, in the relative loudness or softness of the notes.

If you were too sophisticated for this simple exercise, I might ask her to play the little tune from Haydn on which both Handel and Brahms wrote variations. I should then ask her to play the basic variation by each composer, tell you who wrote it, and ask you to tell me the chief difference between the two styles out of four or five suggested differences. Then I would have her play perhaps ten or twelve variations in a random order: some by Handel, some by Brahms, some by neither, and some on the wrong tune. You can see how this sort of thing can get quite complex. If I started you analyzing fugues and telling me what the different voices were doing, I should soon have you hanging on the ropes.

If there is time, as there is in a course, although not in an examination, I like to make a point about any great work in sonata form by playing a complete sonata, quartet, or symphony and substituting one movement that does not belong: that comes out of a similar work by the same composer, written at about the same period of his development. One might think that when a composer writes a work in four movements, he writes one tune for the first movement, another for the second, another for the third, and another for the fourth, and then puts the four tunes together, with perhaps some relationship in key or mood. I am convinced that nothing like this happens; he writes basically the same tune in four different ways—although sometimes, I admit, it takes a good deal of subtlety to recognize it as the same tune. What any sensitive musician can do at once, however, is to recognize it if you put in the wrong minuet, the wrong slow movement, or the wrong finale—and not because he has any previous acquaintance with the work, but simply because they don't belong; they won't fit. I have found this sort of exercise very good for demonstrating the underlying unity of a classic composition in four movements.

For the visual arts, one of the simplest techniques is to use four projectors that can throw four slides at once on a large screen and ask a series of questions about them that can be answered with the number of the correct slide, or with "None of them." For example, at the humblest level: which one was a watercolor? Which was an oil? Which was a fresco? Which was a wash drawing? If they represent well-marked historical styles, which was Italian of the High Renaissance? Which was Dutch? Which was eighteenth century French? Which was eighteenth century English? All such works, of course, must be previously unknown to the students. At a slightly deeper level, one might show three works by one painter and one by a contemporary in a different style, such as three by Cezanne and one by an Impressionist. The painting that does not belong with the others will stick out like a sore thumb to a sensitive student, but others will see no difference. One can ask all sorts of questions about techniques and about the composition. for example, in which is the basic form of the composition a rectangle? a pyramid? a diagonal? an S-curve? Which makes the most obvious use of contrasting textures? Which one is basically two-dimensional? Which is organized in a series of planes? There is hardly any limit to the number of questions one can ask in the form, "Which painting does A, B, C, and D?" If the paintings have clear-cut differences, and if the questions are perceptive, the answers will reveal a good deal of sensitivity to what is going on in a work of art.

I should next like to turn to the field of history as it relates to the humanities, hoping to leave a clear field for Max Engelhart to deal with history as it relates to the social sciences. You will see the difference, I hope, in the kind of illustration I will use. The sort of outcome I have in mind has no large social significance; it is not a necessary ingredient of good citizenship; it has, rather, a personal significance, and is an ingredient in one's philosophy. Less pompously, it is a part of that general stock of ideas that a man carries around in his head that determine what objects and events in everyday life will strike him as significant, or interesting, or puzzling, or dangerous, or good. Among these ideas I would place a high value on a sense of the past and of its continuing influence on the present.

For example, I once read in a book of popular scientific essays the statement that the average person must take the existence of a man like Julius Caesar on faith, or on authority. To a man educated in the humanities, that statement is preposterous. I am sure that at least two-thirds of you have in your pocket or in your purse some tangible evidence that Julius Caesar existed and left us something that we use every day. It is something that you can take out and hold in your hand. It even has the name, Julius, printed on it, although in an English form that you may not at first recognize. The place in which his name is printed is peculiarly appropriate in view of his life history. What is it?

I could nudge you closer to the answer by writing the names of our last four months on a blackboard: September, October, November, December. Then I would erase the "ber," on the ground that it is nothing but a shiver, and would have left four good Latin words: *septem*, *octo*, *novem*, and *decem*. We have other forms of these same words in *septet*, *octet*, *novena*, and *decimal*. What do they mean? Obviously, seven, eight, nine, ten. But why do we call our ninth month the seventh, our tenth month the eighth, and so on? Did somebody lose count? That is hardly likely. We can get a clue as to what probably happened by writing out the Latin names of the months immediately preceding these four: *Julius* and *Augustus mensis*, our July and August. Do the names *Julius* and *Augustus* ring a bell? Certainly: Julius and Augustus *Caesar*. But how did their names happen to get into our calendar and displace all the following months by two months? At this point I allow a little time for research. Someone usually looks up the encyclopedia article on "Calendar" and discovers that, when Julius Caesar was campaigning in Egypt in 48 and 47 B.C., he was not too preoccupied with the charms of Cleopatra to notice that the Egyptians had a much better calendar than the Romans. Consequently, when he returned to Rome in 46, he brought along not only Cleopatra but an Egyptian astronomer named Sesogines, and with his help worked out essentially the calendar we use today. His successor, Augustus, secured its adoption throughout the Roman Empire, and either he or his followers apparently saw to it that the names Julius and Augustus were forever enshrined in it, disregarding the protests of the mathematicians that the names of the last four months would all be two numbers off from their proper numbers. Augustus may have explained that it had to be that way because Julius was born during the month we now call *July* in his honor—and by a surgical procedure that we still call a "Caesarean section."

If I wanted to make this exercise just a bit more complicated, I might ask you why we called this calendar the "Julian Calendar" down to about the time of George Washington in English-speaking countries, and still later in countries dominated by the Greek Orthodox Church, but then shifted over to a slightly modified calendar that we call the "Gregorian Calendar." One of the chief differences between them is that our present dates are eleven days off the corresponding Julian dates. If you look in old history books, you will see that George Washington was born on February 11 (Old Style) or on February 22 (New Style), obviously the Julian and Gregorian dates. One of the best exercises in semantics I know is to have a class argue about the date on which he was "really" born. If we accepted the Julian date, it would make quite a difference in the time of the annual meeting of the American Association of School Administrators. Why did we change? What was done to prevent losing eleven days again? Did it work?

A still more complicated problem is that all our names for the months are Roman, while all our names for the days are Anglo-Saxon, yet each of our days corresponds planet for planet with the Roman names. *Solis dies* and *lunae dies*, day of the sun and day of the moon, Sunday and Monday, are obvious examples. Wednesday, which is Woden's day, is less obvious, but his planet was Mercury, as you can see in the Latin *Mercuri dies* or the French *Mercredi*. Thursday is obviously Thor's day; he was "the thunderer," as you can see in the German name for this day, *Donnerstag*. But so was Jupiter the thunderer; hence this day was *Jovis dies*, the day of Jove, in Latin, or *Jeudi* in French. All the other day names are Anglo-Saxon equivalents of the gods or planets for whom the Romans named their days. When do you think the Anglo-Saxons had a chance to pick up the Roman names for the days and then translate them into their own language? They did not invade England until the Roman Occupation had ended, and they had little if any contact with the Romans on the continent. And why didn't they translate the months?

If you are wondering how I would put problems like these into multiple-choice form, the truth is that I wouldn't bother. I'd give my classes some linguistic data, some historical data, a little time for research, ask them a series of questions, and let them answer in essay form, perhaps after some preliminary discussion. You may be disappointed because you know that the grading of a single essay is unreliable as a test or examination. But I am not talking about a test; I am talking about exercises for a course; and after I get sixteen to twenty such exercises over a period of a year, and grade them the way I grade them, the reliability of the cumulative total score can easily exceed .9.

Please note that the program does not require me to talk exclusively about multiple-choice tests: the topic is "Exercise-Writing."

I might carry this same line of thinking into a somewhat different outcome of the humanities that I call a sense of the interconnection of ideas, especially across languages and cultures. For example, when I learned that the new tranquilizing drugs had been given the ugly name, *ataraxics*, I greeted it as an old friend. It was the ringing battle-cry of my fighting days in college, the watchword of the Epicureans, *Ataraxía*, which may be translated, "Do not be disturbed!" Why it had such an appeal for young men I have never figured out. But I found the same idea expressed in Horace's motto: "Nil admirari," which may be translated, "Don't be swept off your feet!" There is even a hint of it in the Biblical, "Take no thought for the morrow," but that is really a different idea: it means trusting in God, not in your own inner resources. But later on the same idea is picked up in Castiglione's ideal for a courtier, which he called *Sprezzatura*, and in the seventeenth century French ideal of the *honnête homme*. Coming down to our own time, I think we have the same idea on a somewhat different plane in Oxford reserve and Harvard indifference. It may be something basic in the code of the gentleman that has perennial appeal for youth. Our young men in one of their moods seem to be groping for "the still point of the turning world."

Here, again, I would not attempt to put such an exercise into objective form unless I had some compelling reason, such as having to give a test to 50,000 candidates. For a class it would be better just to give a list of these slogans, perhaps with translations, and ask what they had in common, why they have such an appeal for youth, and which one is farthest out of line with the others, as I believe the Biblical passage to be.

But I had better give some attention to objective exercises, and for that purpose I shall choose a poem, because nowhere else can I find so much meaning in such compact form. This one is called *Spring and Fall*, and was written by Gerard Manley Hopkins. It appears to be addressed by a mature man to a young girl, and hence the title might be taken to refer to youth and age. I hope to convince you before we are through that this is definitely an error. *Spring* may well refer to youth, but *fall* in this title has a much more sinister meaning than "age." It is a moot point whether one can ask objective questions about a poem without lacerating it. I believe that one can elicit a very sensitive reading of a poem by objective questions. In fact, when anyone asks me whether I really understand a very difficult poem, like one of Eliot's *Four Quartets*, I have to answer, "I don't know. I haven't yet got around to making up a test on it."

It might be better to begin by reading the poem straight through, but it is rather puzzling and would not "get across" at first reading. Hence we shall take it by pairs of lines and ask somewhat abbreviated objective questions about them.

"Margaret, are you grieving
Over Goldengrove unleaving?"

Here are two puzzles. What is unleaving? Staying? Failing to produce leaves? Unfolding leaves from buds? Or shedding leaves? The only one of these that might cause a young child to grieve is shedding leaves—in autumn. Hence Goldengrove can hardly be a particular flower or shrub but just a patch of woods in autumn, with its characteristic golden color.

"Leaves, like the things of man, you
With your fresh thoughts care for, can you?"

Here the syntax is a bit tangled for the sake of the rhyme, but it obviously means, "Can you, with your fresh thoughts, care for leaves, like the things of man?" We should get at the ability to unravel the syntax by a grammatical question. What would be the *opposite* of "the things of man"? The things of woman? The things of children? Abstract ideas? Probably it is "the things of nature," because what she is caring for at this time is the falling leaves, and the poet expresses some surprise that she can care for these things, which are certainly things of nature, just as she cares for the things of man.

"Ah, as the heart grows older
It will come to such sights colder
By and by, nor spare a sigh
Though worlds of wanwood leafmeal lie."

The puzzle here is the interesting formation, "leafmeal," which I have never seen anywhere else. We have to figure it out by analogy. Is it like oatmeal, cornmeal, last meal, or piecemeal? The first three would be ridiculous but the last suggests a clue. "Piecemeal" means "piece by piece." Can "leafmeal" then mean "leaf by leaf?" It certainly makes sense, for that is the way the leaves would be lying. The word "wanwood" is not in any dictionary, British or American, which is small enough to lie on my desk, and it is hardly worth tracking down in anything larger or more specialized since it obviously has to be some kind of dead foliage which is lying in great quantities, leaf by leaf, upon the ground. As Margaret grows older, she will not spare a sigh at such a sight.

"And yet you *will* weep *and* know why."

As the poem is usually printed, the poet has indicated a strong stress on *will* and on *and* by accent marks. This is a first-rate puzzle, since the poet has just said that she will come to such sights colder and will not spare them so much as a sigh. If she then breaks down and weeps, with no reason given for the change in mood, there is a flat contradiction.

The only way out of it that I can see is that she is weeping now, over the falling of leaves; the poet has tried to comfort her by telling her that by and by she won't care; but she refuses to be comforted: she goes right on weeping and wants to know why she is weeping. This is the *volitional* use of *will*, as in "He *will* do it no matter what you say." It might be translated, "In spite of what I have been saying, you insist on weeping and on knowing why." I can't see any way in which it could be weeping in the future, for he has just said that in the future she not only won't weep; she won't even sigh. Now he begins to tell her why.

"Now no matter, child, the name:
Sorrow's springs are the same."

He doesn't want to tell her just the *name* of what she is weeping for, since now it would probably mean nothing to her; and anyway, the ultimate source of all sorrow is the same ("sorrow's springs").

"Nor mouth had, no, nor mind, expressed
What heart heard of, ghost guessed:"

We have to do a bit of translating here. "Your heart heard of, and your spirit guessed, what had not been stated in words or even formulated as an idea." I should simply ask, "Which of the following is the best reading of lines 12-13?"

"It is the blight man was born for,
It is Margaret you mourn for."

Here is the ultimate answer to the question of why the young girl is weeping. In the falling of the leaves she has unconsciously glimpsed a symbol of her own death, far off in the future: "the blight man was born for." It is really not the falling leaves she mourns for, but for herself, even though it is only what her heart told her, and her spirit guessed; it had never been explained to her in words, nor even entered her mind as an idea. That is why sorrow's springs are the same, for the ultimate source of all sorrow is death. It is also why the name is no matter, now, for it would mean nothing to her, but in her heart she has already guessed it.

If anyone now wants to interpret the title, "Spring and Fall," as "Youth and Age," he must have a sentimental aversion to the obvious interpretation, once we have puzzled it out, which is "Youth and Death"—the latter symbolized by the falling leaves. There are very few interpretations of a cryptic poem that one can absolutely rule out as untenable, but I believe "age" in the title of this poem is one of them. There is no clear reference to old age anywhere in the poem. True, as her heart grows older, she will no longer weep at the falling of leaves, but that could well occur by the age of twelve. The poet is represented as somewhat older than the girl, but there is no reason to suppose that he is very old: he might be a man of thirty. The girl is certainly not weeping about *his* age: "It is the blight man was born for,/It is Margaret you mourn for." That seems to me an unmistakable reference to death.

I have now used up my allotted number of pages and have discussed only five topics: a few techniques for getting at sensitivity to what is going on in a musical composition or a work of art; a few informal ways of revealing understandings that are peculiarly characteristic of the humanities: namely, a sense of the past, and a sense of the interconnection of ideas; and, finally, some objective means of eliciting a sensitive response to a philosophical poem. Since I would use the same types of items to reveal more than a superficial grasp of any prose passage, literary, philosophical, or historical, I hope that in some way or other I have touched upon the chief fields usually associated with the humanities. To attempt a comprehensive coverage of such a vast domain, so differently conceived and differently taught in different institutions, would obviously be absurd in a brief talk. I can only hope that I have left you with a few testing ideas that are fairly representative of the spirit of this great domain.

Exercise Writing in the Natural Sciences

LEO NEDELSKY

This paper is concerned with general principles to be used in writing test exercises, especially objective test exercises, in natural sciences. Most examples will be drawn from physics, the field in which the author feels most at home; the generalizations which these examples are to illustrate, however, should be applicable to any of the natural sciences.

Exercise Variety

In these relatively enlightened times it may be safe to assume that the test writer has before him a list of clearly stated objectives, i.e., a description of the kinds of knowledge and abilities the test is to measure. It is well, however, to look into the origin of the list. The most important function of a science test, perhaps its only defensible function, is to predict the student's behavior when faced with a situation in which an understanding of science is useful and important. Such a situation, which we shall call a criterion situation, may occur in the student's academic career, his life as a citizen, or his work as a scientist.

One way to test the student's criterion competence, i.e., competence to deal with a criterion situation, is to present the student with realistic problems, i.e., with situations that closely resemble those he is likely to face in the future. A test of this "synthetic" sort is quite valid but also almost prohibitively cumbersome and expensive, for genuine problems facing a scientist or a citizen are usually very complex and not of the paper-and-pencil type. Further disadvantages of such a test lie in the difficulty of assembling an adequate sample of problems and of communicating about the precise nature of the test and students' "scores."

Another way, is first to analyze the complex competence to solve genuine problems into its constituents and then to test for the more important but still tractable of these. Such an "analytic" test has all the usual shortcomings of an analytic representation of complex and incompletely understood phenomena: neither are all the constituents known nor do they add up to the whole. In addition, some of the more important constituents, such as habits of thought and attitudes, cannot be conveniently or accurately measured. Those shortcomings of an analytic test which result from the unavailability of a complete analysis of a complex competence are probably best alleviated by including in the test exercises that, in the aggregate, evoke in the student a great variety of mental processes in varied patterns.

We are here concerned with the analytic type of test. The list of objectives is the list of presumed constituents of the criterion competence. As has been suggested in the last paragraph, the test writer can in some respects transcend statements of objectives and come closer to the criterion competence by varying test exercises as widely as the limits of the formally stated objective permit. Thus, e.g., in testing for the ability to interpret physics data the following variations are possible. The data may be graphical, numerical, or verbal in form; the content may be, e.g., mechanics or heat; the student may be asked to draw possible conclusions, to assess the accuracy or consistency of the data within themselves or with other data, or to estimate the cogency of the data as evidence for a generalization. The form of the test may also vary: essay or objective.

Exercise Writer's Preparation

If the variety of exercises described in the preceding paragraph is not to be merely haphazard, the exercise writer should not only know the subject matter tested but also have some understanding of what is involved in solving a genuine scientific problem at a level of sophistication higher than that of the students. As will be shown below, he must also know the extent and the range of the students' knowledge, abilities, and general intelligence. He must even know their method of preparation, enough at least to be able to judge the degree of novelty a particular test exercise will have for them. Finally, the exercise writer must have some notion of the more common methods used by the students in solving various exercises.

Objective as Function of the Exercise and of the Students

Having described the prerequisites for writing an exercise we shall skip the actual technique of producing its first version and deal instead with the general methods of criticizing an already written exercise. The key question is of course what a particular exercise measures. This question we shall assume to be equivalent to the following one: If the exercise is successfully performed by a group of students and unsuccessfully by another group, what is the main difference in the abilities of the two groups? It seems clear at the outset that the answer depends on the choice of the whole group. We shall nevertheless labor the point at some length because of the prevalent misconception that the ability measured by a test is a function of the test alone. Some test makers seem to think, for example, that a test designed to pick out promising college material from among the students of an eastern preparatory school can do the same job effectively when given to the students of a rural school in the South. We shall illustrate our point by an example.

A group of students is presented with the following exercise:

It is possible to hear the sound of a fountain behind a brick garden wall although the fountain cannot be seen. This phenomenon
A—can
B—cannot
be explained on the basis that long waves are diffracted more than short waves.

The group will be divided into two subgroups: those who pass by choosing the right response, A, and those who fail. In what respect do these subgroups differ? That is, what relative ability or knowledge does the exercise test for? If the group consists of graduate students of physics of a Spanish University, it seems clear that the two subgroups will differ in their ability to read English, for it may be safely assumed that all members of this group possess the requisite mastery of physics but that only some know English. Next, let us suppose that the group consists of students thoroughly trained in the theory of wave motion, and accustomed to using the terms bending and spreading in place of the term diffraction. The two subgroups are most likely to differ in their knowledge of terminology. Let our next group consist of students who have had a year of physics, a good discussion of diffraction of both light and sound, but no treatment of the relevant similarities between them. Let us assume that most of these students know that sound waves are longer than those of light. The two subgroups are most likely to differ in their ability to surmise that the principles of diffraction are applicable to the situation described in the exercise, an ability which may be considered a part of the more general ability to relate generalizations to specific situations. Let our last group consist of students to whom the particular test situation—sound of a fountain behind a wall—was explained in class or textbook. The ability to recall the situation may be influenced by a variety of factors, especially if the explanation was not emphasized. The main difference between the two subgroups is hard to determine and may not depend on anything educationally significant. An exercise writer who uses situations very similar to those in the textbook even in a test of pure information is on slippery ground.

It is sometimes argued that every pair of subgroups discussed in the preceding paragraphs exhibits the same important difference—the passing group knows the answer to the question while the failing group does not—and that, in measuring the general mastery of physics, it is a matter of secondary importance *how* the various students arrive at the correct answer. It seems that this argument is based on at least two questionable assumptions. One of these is the assumption that it is important to know the answer to the question of the exercise. Surely, unless the student is going to be a landscape gardener or fountain builder, we are interested in his ability to answer not this particular question but this *kind* or class of question, the basis of the classification being the objective to be measured. Let us assume, however, that it *is* important for the student to know the answer to some particular question, for of course there are such questions in science. Granted this, the second false assumption in the argument which we are criticizing is that it is important that the student know the answer on the date of the examination. For if we are interested in the retention of knowledge, in the student's ability to answer the question some months after the examination date, such ability will depend crucially on how the student was able to arrive at the correct answer in the first place.

We freely admit that to make our point clear we have chosen groups that are extreme in their differences. It is nevertheless quite generally true that a given test will measure different abilities if used with different groups. It is the examiner's responsibility reliably to determine what the differences are and consequently what conclusions from test results are justified. The specifications given to the exercise writer must include a description of the group or groups that will take the test.

Necessary and Sufficient Abilities

We have argued above that, in general, the relative scores of the members of a group will depend only on the factors with respect to which the group is heterogeneous. For example, our group number one was homogeneous at the requisite level of mastery of physics, but presumably heterogeneous with respect to the ability to read English. If it is desired to know the relative standing of the members of a group with respect to an ability, the principal ability of the exercise, this ability should clearly be a necessary one for doing the exercise, and the group should be fairly homogeneous with respect to all other auxiliary necessary abilities. The homogeneity of the group relative to auxiliary abilities, that an exercise requires, is best attained by writing the exercise in such a way that the level of the required auxiliary abilities is quite low, i.e., below that of the great majority of the group. If a single grade is to be assigned on an achievement test, it may be sufficient to keep the required auxiliary abilities below the barely passing level. In all cases, however, it is not enough to assert that the students ought to possess the auxiliary abilities; it is necessary to ascertain that the great majority of them in fact do.

In the exercise under consideration the necessary abilities are English, terminology, certain knowledge of sound and light and diffraction, and, finally, the ability to relate this knowledge to a particular situation. Let us now assume that the latter is the ability we want to measure, and that we have a realistic group consisting of students who have had one year of college physics, including the study of diffraction, but who are not likely to have discussed the particular test situation in class. If it seems likely that a sizable fraction of the group, say one-fourth, does not know what diffraction means, the term may be explained or omitted from the exercise. Either of these emendations will make the group properly homogeneous relative to the auxiliary ability of the knowledge of terminology but will make the exercise either longer or less clear to those who do know the term. For such a group it may be better to write an altogether different exercise. The most effective compromise between a reliable measure of an ability and a test of a practical length is determined by the group for which the test is designed.

Besides being necessary, the principal ability of the exercise also ought to be *sufficient* for dealing with the exercise. The following simple example will clarify the meaning of necessity and sufficiency. The student is asked: "What is the product of 3 and 7?" The ability to multiply is not here necessary because the problem can be solved by addition. Nor is it sufficient, for the term "product" must be understood. It should be clear that it is to make the principal ability sufficient or effectively sufficient that the auxiliary abilities must be kept at a low level. If the principal ability is necessary but not sufficient, those who can do the exercise have the ability and those who can't may or may not have it. If the ability is sufficient but not necessary, those who can't do the exercise don't possess the ability and those who can may or may not have it. In achievement testing, both the necessity and sufficiency conditions must be strived for, but often one has to be satisfied with a more modest claim that the ability tested for is likely to be very helpful in dealing with the exercise.

Students' Mental Processes and Validity

There are statistical methods that help us find out just what ability an exercise measures. They all involve correlations between the exercise and a test whose validity is known to be high. Such tests are not usually available, however, for the more complex educational objectives, and the exercise writer must still mainly depend on his surmise of the students' mental activities while working on the exercise. One of the guides toward identifying the measured ability is a formal analysis of the test exercise. If, for example, in an objective item, the term "data" seems to be applicable to the information given the student, and if the responses among which he is to choose can be legitimately characterized as "interpretations" of the data, it is usually assumed that the exercise measures the ability to interpret data. Studies at Chicago (1) indicate, however, that students vary a great deal in the method of arriving at the right response. Some of them read the stem—in our case, the data and the question—arrive at some interpretation, and look among the prepared responses for a similar one. Others test each response in turn against the data and reject the inconsistent ones. If the inconsistencies can be established by inspecting parts of data only, the right response can be chosen without understanding even the general trend of the data. Not only in this example but generally, the abilities required in going from the question to the responses and vice-versa are quite different. It is in general impossible, for example, to use objective exercises for a reliable and consistent differentiation between the old standbys, "Ability to interpret data" and "Ability to apply principles." It is therefore better to use a classification of objectives that depends on the nature of the situation as defined by both the question and the responses and to use "interpretation of data" and "application of principles" exercises only to insure variety (2).

Homogeneity of Responses

Although there are many ways in which students attack objective exercises, in almost all of them the rejection of wrong responses seems to play an important role. If, therefore, the exercise is to measure a specific ability, this ability must be the key one not only in recognizing the right response but also in rejecting the wrong ones. This requires a certain *homogeneity* among the responses. Thus, e.g., if the exercise is to measure the ability to estimate the cogency of facts as evidence for a theory, the responses should all be correctly stated facts and differ among themselves only in their value as evidence. If, on the contrary, the wrong responses contain data that are factually false but that, if true, would have been good evidence, the exercise may or may not measure straight knowledge, but its reliability of measuring the principal ability is reduced.

The "ability-homogeneity" of responses has other values that may transcend that of insuring the purity of the ability measured. We believe that the main advantage of objective test exercises is their ability to define the problem for the student with a precision that is entirely out of reach of essay exercises. The definition must start with the stem, i.e., the part that precedes the responses. The stem must be so suggestive or directive that it guides the student's thinking into proper channels and even enables him to anticipate the general form of the responses. Stems like "Which of the following is true?" may produce tension and start some speculations in the student's mind—perhaps in connection with the preceding item—that are merely distracting. The very first response must further define the problem or, in many cases, even complete its definition. After reading it, the better students should be able to formulate a response that resembles closely in form and substance the right response—it may of course be the first one. And when the student, who has the relevant ability, comes upon the right response, he should recognize it as such and give only casual attention to the following responses. To this end in particular and to increase the general directiveness of the stem, the stem may well include hints as to the nature of the responses. For example, "All of the following responses are factually correct. Choose the one that contains the most convincing (or best established, or most precise) evidence for the theory." Or, "None of the responses below is strictly correct. Choose the one that deviates from the correct response through being too general (or too specific, or quantitatively inaccurate). The problem of communication is a thorny one; every part of the exercise must contribute to the student's understanding of what the problem is and what kind of answer as to form, content, precision, etc. is expected of him. The argument in this paragraph should make it clear why we have avoided using the term "distractor," a term that applies with such damning accuracy to many wrong responses in the existing tests.

Homogeneity of responses also contributes to economy of effort and time. Time allowance for objective items is usually indecently short. Yet it is patently impossible to test the student's thinking ability without allowing him to think. If the student's problem is to evaluate the cogency of evidence, he should be allowed to concentrate on that problem without being sidetracked into estimating the reliability of data or other matters. Nor should he be on the lookout for verbal traps. The "onlys," "nevers," and other words so frequently used to differentiate between the right and wrong responses should be prominently displayed. If underlining such words ruins the exercise, chances are it was not very good to begin with.

Homogeneity of responses relative to the ability measured helps define the problem, concentrates the student's attention on it, and reduces his tension. A similarity in the *form* of the right and the wrong responses, i.e., in their length, presence of qualifying words and phrases, the degree of technicality, and in other essentially superficial respects, should help prevent successful guessing based on the auxiliary ability of test-sophistication. Ideally, responses should differ in nothing but a single quality; in the above example this quality is their cogency as evidence for the theory.

The Best and the Correct Response

In natural science, as in other disciplines, it is nearly impossible to give a strictly correct answer to any but the most trivial question in less than a page. In order to decide how correct the best response should be we must again recall that the main function of an exercise is to divide the students into two groups that differ in a particular ability. The response marked right should then have appeal to those who have this ability. It is therefore at best useless to increase the accuracy of the best response by adding qualifications whose absence would not be noticed by the great majority of the students, say by B+ students and those below. For an obvious example, the relativistic and quantum mechanical corrections or qualifications are out of place in a test over a one-year physics course. It is of course true that the absence of certain qualifications may disturb an exceptionally well-informed student. This is to be preferred, however, to changing the exercise so as to lower its discriminating power for the rest of the class. For, as qualifying phrases are added to make the best response more nearly correct, two difficulties arise. First, a large number of students become puzzled by the presence of the over-refined qualifications, and second, making the formal attractiveness of wrong responses approximately equal that of the best response becomes more difficult. The optimum correctness of the best response must of course vary with the caliber of the students. We are thus reminded once more that the really effective test is tailor-made to fit a particular population.

In some exercises it is preferable not to have the best response correct even in the modest sense of the preceding paragraph. It is then of course usually desirable to warn the students of this fact. By using as the right response one that is not correct but is merely the best of those available, it may be possible to force the student to do more profound thinking. For example, the best response in the following exercise is A.

Exercise: Which of the following is the best definition of *Potential Energy?* (None of the definitions is strictly correct)

A—The energy which a body possesses because of its position.
B—The energy which a body at rest possesses because of its position.
C—The maximum energy which a body can acquire.
D—The energy which a body possesses before it starts doing work.
E—The energy which enables a body to do work.

If the right response were made more correct by including elastic potential energy, its form might easily become so like that in textbooks, that the exercise could be worked by rote memory. As it stands, the exercise is likely to discriminate between those who understand the term, potential energy, and those who don't. It may be remarked in passing that the responses of this exercise lack in formal homogeneity because two of them, A and B, are "paired," i.e., have a close similarity, while the rest of the responses are not. Students soon learn that paired responses are more likely to contain the right one.

Teachers are almost invariably unhappy about incorrect answers even if they are shown good statistical evidence that the exercise correlates highly with the teachers' own choice of good students. Their arguments against incorrect best responses vary from good to bad, the bad ones being more frequent. An example of a good argument is that since we have no accurate knowledge of the mental processes involved in the criterion situation, i.e., in resolving a genuine scientific problem, or those used in the test situation, we should hold firmly to the few similarities between the two situations that are under our control. Thus in a criterion situation of almost any sort it is a statement of the correct definition or law that is useful; therefore similarly correct statements should be used in test situations. There is a good deal of truth in this analysis, although it should be noted that the choice of the best approximation concerns the scientist and the citizen more often than the choice between the correct and the wrong. A less good argument, based on the uncontestable truth that tests are valuable tools of instruction, claims that therefore they should contain nothing but the truth. What the student is likely to learn from a test, however, is not an isolated fact or generalization, for the test is encompassed in a few pages and a few hours as compared to the hundreds of pages of the text studied over a period of months. The most effective role of the test as a teaching instrument is rather to let the student know in an emphatic manner what the objectives of the course are. Another questionable argument runs as follows: "I told my students that 'the energy which a body possesses because of its position' is not a correct definition of potential energy. It is not fair to ask them to accept it." Our reply to this teacher would be that the only physics test that is fair to the students is the one that reliably identifies those of them who understand physics. The question is rather whether the exercise in question is fair or kind to the teacher. On this point we would say first, that fallibility of teachers and texts is not a bad thing for students to learn, and second, that the teacher should have given his class a deeper criticism of the quoted definition than merely calling it wrong. Whatever the worth of the arguments against exercises with incorrect best responses, it should be clear that writing such exercises does require a steady hand.

The Wrong Responses

The prevalent criterion for a wrong response seems to be that it should be wrong and yet plausible. As we have indicated before, this requirement, although necessary, is not enough. If the student is asked to choose among conclusions from some data, the wrong responses should be wrong only in their relation to the data. Of course even if the wrong responses are homogeneous in this respect, there are still degrees and kinds of wrongness that are possible. The degree of wrongness determines the difficulty of the item and sets the main discrimination line, e.g., between A and B or between D and F students. The *kind* of wrongness determines the auxiliary abilities that are helpful in finding the right response. Thus, e.g., if the wrongness in the preceding example is that of going beyond data, students who have such a weakness will be more attracted to the wrong responses than, say, students who have the weakness of being over-cautious. Besides these two, there are many other weaknesses that make wrong responses and sometimes the right response, unequally attractive to students who possess the principal ability in an equal degree. Two common ones are distrust of theory and over-reliance on quantitative data. A play on various students' weaknesses to make wrong responses attractive to them is often justified and, in fact, can seldom be avoided. It is necessary, however, that the prevalent weaknesses be sampled fairly. Such sampling is controlled primarily by the nature of the wrong responses and to a smaller degree of the right response if it deviates, as it usually must, from the strictly correct one.

Although it is difficult, and unnecessary, to have a single exercise discriminate at several levels of ability, it is quite easy and useful to make it discriminate at two levels by making one or two of the wrong responses so wrong that only failing students should be attracted by it. Let us call such a response an F-response, and the right response, R-response. Let the number of F-responses a student chooses be called his F-score, and similarly for R. It will almost always be found that the R-minus-F scores are more reliable than the R-scores; F-scores may discriminate more reliably between F-students and the rest than any other scores. It should be noted that for students who make a try at every exercise, the reliability of the usual right-minus-wrong scores is identical with that of the R-scores. F-scores are also useful for establishing absolute standards for passing performance (3).

Spontaneity

If the strongest argument for objective tests is that they make it possible to make it clear to the student what is expected of him, the strongest argument for the essay test—we are not here considering the ability to express oneself—is that it calls for a spontaneously produced answer. Since all criterion situations call for greater spontaneity than is required in the usual objective test, the latter, if not supplemented by an essay test, must be modified to decrease such discrepancy in some of the exercises. An objective exercise that comes closest to this requirement is one in which the student reads the statement of the problem, decides on the right answer, and then searches for it among the prepared responses. Such a procedure can never be enforced but it can be made more profitable than any other. To this end the stem must be very directive and the responses non-directive or even non-evocative, i.e., non-suggestive.

The following exercise has non-evocative responses. The student is asked to solve two simultaneous equations: $3x+2y=17$, and $x-y=9$. If the responses were pairs of values of x and y, some students would find it easier not to solve the equations but rather to substitute the pairs of values into the equations. If, on the other hand, the correct response is given as $x+y=5$, and the wrong ones similarly, a solution of two simultaneous equations is the only way to the right answer, and solving the given two is by far the quickest way.

Numerical science problems lend themselves well to the technique of non-evocative responses. For example, each response may list just the second digit of a numerical answer. Or, the student may be asked to compute two quantities, or two formulas, and asked to choose among responses that are ratios of the two. In qualitative exercises, a similar result may be achieved by making the analysis of each response, without first solving the problem of the exercise, as much of a chore as solving the problem. The following exercise, in which, to save space, we show only two responses, may serve as an illustration.

Exercise: This exercise involves two steps. First, decide what law of nature is most directly useful in explaining the following fact: A brick can be pulled along a fairly smooth surface by means of a string; the string would break, however, if jerked sharply. Second, choose that one of the following phenomena for which this law of nature provides an explanation.

A—A glass tube dropped from a height of 10 feet breaks if it falls on a concrete sidewalk but will not break if it falls on a soft ground.

B—It is impossible to lift oneself by pulling up on one's own hair.

Most students will find that the most efficient way to deal with the above exercise is to follow the directions which suggest giving a spontaneous answer to the first question. Directions to the students must of course always be scrupulously honest in the sense of indicating the easiest path to the right response. We note that the second part of the stem in the above exercise is not at all directive. This is the usual price we must pay for introducing spontaneity into objective exercises. A more elaborate but also easier method to combine the more advantageous aspects of the essay and objective exercises is described in another article (4).

Summary

An effective test must be tailor-made for a particular population. The degree and range of the students' abilities determine: the optimum correctness of the best response; the degree and kind of wrongness of the wrong responses; the level of the auxiliary abilities necessary for working the exercise. This level should be below that of the great majority of the students.

The definition of the ability measured by an exercise should hold true for all the prevalent methods of solution used by students; e.g., the method of eliminating wrong responses.

The principal ability of the exercise should be both necessary and sufficient for working the exercise.

Test exercises measuring a particular ability should exhibit as great variety as is consonant with the definition of the ability. They can vary in content, form, type of analysis required, difficulty, and auxiliary abilities.

The situation on which an exercise is based should not be similar to the textbook ones.

Abilities auxiliary to a group of items should be sampled fairly by these.

The stem of an exercise should contain a clear statement of the problem and even suggest the desired kind of solution.

Responses should be homogeneous relative to the principal ability and in their form.

The technique of using incorrect best responses is useful but difficult.

It is easy and desirable to include among wrong responses some that are attractive to failing students only.

It is possible to make an objective exercise evoke in the students nearly spontaneous answers by making this process the most economical one for dealing with the exercise.

REFERENCES

1. *Problem-solving Processes of College Students*, by BENJAMIN S. BLOOM AND LOIS J. BRODER, University of Chicago Press, July 1950.
2. "Formulation of Objectives of Teaching in the Physical Sciences," by LEO NEDELSKY, *American Journal of Physics*, 17, p. 345, September 1949.
3. See the author's articles in *Educational and Psychological Measurement*; "Absolute Grading Standards for Objective Tests," Spring, 1954, p. 3, and "Ability to Avoid Gross Error as a Measure of Achievement," Autumn, 1954, p. 459.
4. "Evaluation of Essays by Objective Tests," by LEO NEDELSKY, *Journal of General Education*, VII, p. 209, April 1953.

Exercise Writing in the Social Sciences

MAX D. ENGELHART

In discussing the art of exercise writing in the social sciences, it seems wise to limit the discussion to exercises useful in evaluating certain important objectives of a college level general course in social science. Many of the suggestions made also have application to more specialized courses in the social science field on both the high school and college levels. Before considering the writing of exercises, it is desirable to characterize briefly what constitute, in my judgment, desirable instructional objectives, methods of instruction, and subject-matter content of a social science general course.

Instructional Objectives in Social Science

As in other subjects, the instructional objectives of such a course may be classified under the headings of (1) knowledge, (2) intellectual skills, and (3) ideals, attitudes, interests, and appreciations. Under the first heading can be listed knowledge or understanding of specific facts, terminology, and principles. We may also include knowledge of the methods of inquiry in social science and how social science can contribute to the making of choices between values, without determining, as science, which values to choose.

Under intellectual skills, we may include the skills required in making discriminations and comparisons and in organizing knowledge in ways which contribute to understanding of relationships. We may include the skills required in reading discussions of social problems or issues with comprehension and with sensitivity to the logic, or lack of logic, of what is read, presuming that these discussions are of varying points of view, or are based on evidence of varying relevance and dependability. Under intellectual skills we may include the ability to analyze a social problem, to recognize assumptions and to propose hypotheses, to obtain relevant data from appropriate sources and to arrive at warranted conclusions. Especially important is the acquisition of the ability to predict and to compare the consequences of different courses of action. This objective is fundamental to the critical evaluation of social policies essential to effective citizenship.

There are, of course, no sharp boundaries between the levels of objectives just mentioned. This is also true of the classification of objectives in the *Taxonomy of Educational Objectives* by Bloom and others.(2)* Furthermore, an exercise may evaluate the attainment of different specific objectives given different learning experiences.

We shall not here be concerned with the formulating of essay questions nor with the evaluation of ideals, attitudes, interests, and appreciations. In the bibliography of this paper the speaker has listed the provocative article of V. M. Sims entitled "The Essay Question is a Projective Technique" and John Stalnaker's scholarly chapter on essay examinations in Lindquist, *et al.*, *Educational Measurement.* (10) A comprehensive program of measurement in a social science general course should include some amount of essay testing, the writing of papers on social problems, and the evaluation of change in social attitudes and beliefs. Levi's *General Education in the Social Studies* (8) is one source of information with respect to such evaluation.

The Content and Methods of Social Science Instruction

Where teaching is restricted to the imparting of the content of a text, in fairness to the students, evaluation should be restricted to measurement of the information they have thus obtained. On the other hand instruction which justifies the use of exercises measuring intellectual skills should provide for the development of such skills. In addition to the explicit recognition of these skills as objectives, instruction should give students numerous opportunities to acquire and practice them. In class discussion and in written assignments, thought-provoking problems should frequently be presented. The instructor should be constructively critical of the thinking done by students whether in recitation or in written work. While much time must be devoted to the imparting of knowledge by an instructor and to its acquisition by students, instruction can include development of understanding of the nature of assumptions and hypotheses, and of the elements of logical reasoning. Students should be expected to use terms with precision and to support their answers with evidence. Problems and issues should be analyzed in class so that students may be trained in identifying problems or issues, in recognizing assumptions, in determining what kinds of data are needed to support or disprove hypotheses, and what methods are useful in collecting and in interpreting the data. From time to time instruction should include the comparing and contrasting of facts, ideas, principles, and generalizations earlier learned with those being learned so that the student is able to organize his knowledge in ways which will prove helpful in solving new problems.

---

*The numbers in parentheses refer to the numbered items in the references at the end of the paper.

While there are numerous sources of information concerning the methodology of social science instruction, reference will be made only to two recent and challenging books—*Theory and Practice of the Social Studies* by Earl Johnson (6) and *Teaching High School Social Studies* by Maurice Hunt and Lawrence Metcalf. (5) The latter should be of equal interest to teachers of the social studies on the college level. It is unique in its applications of field psychology to learning in social science. It strongly advocates the use of content relevant to areas of conflict and of contradictory beliefs both within persons and between persons such as social class; race and minority group relations; sex, courtship, and marriage; and religion and morality. Whether or not these particular areas are included, the content of the social science general course should include areas in which there are important controversial problems and issues. *Contemporary Social Issues* by Lee, Burkhart, and Shaw, (7) and *Basic Issues of American Democracy* by Bishop and Hendel (1) are examples of sources in which students may read selections presenting opposing points of view on numerous problems and issues. *Society and Man* by Weinberg and Shabat (11) is unique in that its authors have shortened and rewritten, in language understandable to students, but acceptable to the original authors, research studies and basic writings of noted authorities in the social science field.

Unless the students have had previous instruction relevant to fundamental concepts and principles of sociology, economics, and government, it is possibly unwise to base a general course on such a "problems" text alone. There may be, however, concurrent use of a systematic text and a problems text. While instruction and class discussion should be concerned with problems and issues, it is more effective to deal intensively with a relatively few problems or issues in relation to more systematic content and in relation to social science methods of inquiry, than to deal with numerous problems or issues superficially. With all of the preceding as introduction, let us turn to the art of exercise writing in the social sciences.

The Role of the Teacher in Exercise Writing

While it would seem logical to insist that exercise writing begin with a carefully formulated and detailed list of specific objectives and a detailed analysis of the subject matter content to be covered, it is my experience that such a formal approach is seldom effective with a group of teachers. It should be emphasized, however, that a less formal approach with the underlying goal of gradual development toward adequate definition of objectives and representative sampling of content can lead, after a number of semesters, to the production of exercises compatible with a wide range of objectives. There will be an increasing emphasis on intellectual skills or critical thinking in the social science field both in instruction and in evaluation of the goals of instruction.

After numerous exercises have been written by the various teachers of a group, its members should critically evaluate each other's work. When different instructors key exercises independently, discussion of disagreements with respect to correct answers can stimulate improvement of the content and phraseology of the exercises. Group evaluation of exercises may result in the elimination of exercises relevant to trivial content or to content uniquely taught by one of the teachers. Group evaluation may also result in the organization of an examination representative of the content of the course and of the objectives it is desired to evaluate. After the examination has been given and subjected to item analysis, the item difficulties, expressed as per cents of correct response, make it possible for each teacher to evaluate achievement in terms of specific objectives. The item difficulties and the item-test correlations are also of great help to teachers in revising series of exercises for future use. It is especially true of social science exercises that knowledge of the per cents of response to each answer, both correct and incorrect, contributes significantly to the improvement of faulty exercises.

The Writing of Multiple-Choice Exercises

As in other fields the writer of multiple-choice exercises in social science should observe a number of precautions. The introductory part of each exercise, the "item stem" should present the problem of the exercise. This promotes both clearer understanding of the problem of the exercise and the use of briefer answers, which often may be no more than single words or phrases. Such exercises are more economical of testing time and since more exercises can be used the total test may be more valid and reliable. Unless the exercise begins with a question, each answer should complete the item stem grammatically. One of the answers should be definitely correct and the other answers should be plausible although incorrect. (In exercises of the best-answer type one of the answers should definitely be the best.) Frequently incorrect answers, or "distractors" can be true in themselves, but not relevant to the problem set by the item stem. Often an exercise writer can produce good distractors by anticipating the kinds of answers students are likely to select when applying the wrong information or the wrong kind of thinking to the problem. Precision in the use of English and the ability to express complex ideas in brief and simple phraseology are desirable attributes of an exercise writer. Some of the suggestions just made may be illustrated by quoting two multiple-choice exercises.

1. Authorities on marriage deplore the American "idealistic belief in romantic love." A young couple characterized by this belief will be more likely to have a successful and happy marriage if also characterized by
   A. a belief in the double standard of morality.
   B. a willingness to take risks.
   C. emotional maturity and adaptability.
   D. a belief in the contractual concept of marriage.
   E. physical attractiveness.
2. A large increase in total spending by consumers, business, industry, and government will probably not cause inflation if
   A. government bonds are sold only to individuals.
   B. interest rates on loans are reduced.
   C. production of goods increases in proportion to spending.
   D. less money is saved by individuals.
   E. money is placed in circulation as needed for spending.

Classification or Key-List Exercises

While many and sometimes all of the exercises in a social science examination may be of the multiple-choice type and "self-contained" in the sense that a given exercise is not a member of a series of related exercises, it is frequently effective to use series of items of the classification or key-list types. In my judgment, items of these types can be useful in evaluating understanding of relationships and ability to make comparisons and discriminations. The expectation of such series of items, ranging more widely over content than self-contained multiple-choice exercises, can motivate students to organize their knowledge, an advantage sometimes claimed for essay testing by persons critical of objective tests.

Suppose, for example, that in a social science course, instruction has been concerned with the characteristics of liberal democracy, communism, and fascism, and that the instruction has included some discussion involving the contrasting and comparing of these ideologies. Then a series of exercises such as the following is appropriate.

After each item number on the answer sheet, blacken *one* lettered space to designate that the item is characteristic of the theory of

A liberal democracy.
B Communism.
C Fascism.
D both Communism and Fascism.
E both liberal democracy and Communism.

1. There should be respect for individual personality in both the ends and means of government.
2. Freedom and equality are meaningful only in a classless society.
3. A temporary dictatorship may precede the establishment of a "stateless" society.
4. The ultimate goal is freedom and equality for all in a democratic society.

Recall of the comparisons made during instruction may enable many students to classify such items correctly. Something more than mere recall is needed, however, if the items are not in the same words used by the instructor and if the student has to select the ideas that are relevant from a wide range of information. In writing such exercises a number of precautions should be observed. The categories should be related, but mutually exclusive. If, for example, the categories pertaining to liberal democracy, communism, and fascism had included socialism it would be difficult to classify such an item as "Advocates collective ownership of the means of production." Careful wording of the categories is extremely important. The word "theory" was included in the directions since it was evident to the writer of this series that the students could become confused if they thought of practices rather than theories. The items should not be long, involved, and qualified complex sentences.

In thinking critically about some controversial problem or issue in the social field, students need to become alert to arguments that a protagonist for one side of the issue will use to support his side and the arguments he will advance against the other side. Consider in this connection the following key-list categories and a few of the items to be classified in accordance with them.

Imagine two persons debating the relative merits and limitations of presidential and parliamentary governments. After each item number on the answer sheet, blacken *one* lettered space to designate that the item is an argument advanced

A in support of presidential government.
B in opposition to presidential government.
C in support of parliamentary government.
D in opposition to parliamentary government.

1. Such a government is especially designed to prevent a dangerous concentration of power.
2. Such a government should respond more quickly to the expressed will of the people.
3. Experience shows that a government which permits the executive and the legislative to become rivals for power retards governmental action.

4. Such a government is especially ineffective where there are several political parties.

Exercises Relevant to Quoted Material

In evaluating intellectual skills, it is effective to use series of multiple-choice exercises or key-list items relevant to brief selections quoted within the examination itself, or assigned to students for critical reading prior to the examination. Selections presenting different points of view or advocating different courses of action are particularly useful. On occasion, exercises following a brief selection may call for student analysis of its content in relation to that of one or more other selections earlier assigned. In a recent examination, a paragraph from *American Community Behavior* by Jessie Bernard attributes minority group conflict to competition for jobs and scarce consumer goods including housing. Three of the multiple-choice exercises following this paragraph expect the student to compare its thesis with the points of view on minority group relations expressed by Robert Redfield, T. V. Smith, and Gunnar Myrdal in selections earlier studied. One of the exercises follows:

1. In his *American Dilemma*, Myrdal disagrees with the above writer in that Myrdal
   A. opposes racial discrimination.
   B. explains discrimination in terms of inconsistent social values.
   C. ignores the role of economic factors in causing racial prejudice.
   D. favors a certain degree of inequality as being socially healthy.

Each answer in the following exercise is quoted from a paragraph from Jefferson's first inaugural address presented above the exercises.

1. In view of Hamilton's beliefs with respect to our government, it is likely that he would consider least desirable
   A. "Equal and exact justice to all men."
   B. "The support of the state governments in all their rights."
   C. "The supremacy of the civil over the military authority."
   D. "The honest payment of our debts."

When two or three brief selections advocating contrasting points of view or differing courses of action are quoted within an examination they may be followed by multiple-choice exercises or by key-list items. For example, one selection may be a quotation from Edmund Burke's "Address to the Electors of Bristol" while the other may be a paragraph by Speaker Rayburn on the importance to a legislator of respecting the wishes of his constituents. A series of items may follow such directions as:

For each of the following items, blacken *one* lettered space if the item is one with which

A Burke would agree.
B Rayburn would agree.
C both would agree.
D neither would agree.

Certain things need to be considered in selecting material for quotation within a social science examination. In addition to relevance to instruction and elements of novelty, the quoted material should not be of unreasonable length. It should say much in little space. In some cases it is necessary to adapt rather than to quote literally. Often it is desirable to substitute words or phrases more readily understood by students. On one occasion, "the architectural drawing of a benevolent welfare state" was changed to "the idea of a welfare state." Frequently, modification of the quotation may be guided by the exercises written. Sentences which do not contribute to the solution of any exercises may be omitted if this does not impair the major thought of the selection. Sometimes an exercise can be keyed more readily if a word or phrase is modified. When a quotation has been changed as described above the citation of its source should begin "Adapted from . . . ."

Some of the exercises pertaining to quoted selections may evaluate "background" knowledge of facts, terminology, principles, or conditions relevant to the quoted material, but not defined or explained therein. While certain of the exercises may be written to evaluate knowledge, other exercises may be written to evaluate the acquisition of the intellectual skills characteristic of critical thinking. For example, an exercise may call for the identification of a central issue or problem:

1. The basic problem of all government involved in the situation described above is
   A. free enterprise vs. socialism
   B. democracy vs. dictatorship.
   C. individual freedom vs. the welfare of society.
   D. government by an elite vs. government by the masses.

Another exercise may have to do with the identification of an assumption or an hypothesis. For example:

1. The above paragraph describes how interviewers collect data to be used in predicting which party will win an election. The *basic* assumption made is that
   A. the persons interviewed have sound reasons for voting as they plan to do.

B. the persons interviewed are well informed with respect to the merits of the candidates.
C. all age groups in the population are represented.
D. the persons interviewed are representative of all the voters on election day.

Examples are given below of exercises calling for the identification in a quoted selection of bias, prejudice, and propaganda devices; of inconsistencies in an argument; and of limitations in the data presented, or in the techniques used in collecting them.

1. In describing Mr. Brown's argument, one can justifiably say that
   A. he was consistently factual.
   B. he was logical and precise.
   C. he used propaganda devices.
   D. he discussed both sides of the issue.
2. Although Mr. Brown is for "free enterprise," he is inconsistent in asking for
   A. an elimination of the excess profits tax.
   B. reduction of government expenditures.
   C. elimination of subsidies for farmers.
   D. government insurance of home loans made by private agencies.
3. The data reported in the table show that a greater proportion of children from broken homes become delinquent than other children. The importance of this factor could be better evaluated if we knew
   A. the comparative intelligence of delinquent and non-delinquent children.
   B. the economic status of the homes of both types of children.
   C. the educational opportunities of both types of children.
   D. all of the above.
4. The author of the experiment summarized above concludes that children learn more efficiently when motivated by praise rather than reproof. We would be more confident of the results of this experiment had the experimenter
   A. had the pupils alternately praised and reproved during instruction.
   B. used a control group of pupils of equivalent initial ability subjected to reproof.
   C. used an equivalent control group neither praised nor reproved.
   D. tested a different hypothesis.

Certain exercises may require discrimination between expressions of fact, opinion, or value judgments:

1. Instead of stating a fact, the sentence in the paragraph which expresses a belief or opinion begins
   A. The price of agricultural products . . .
   B. The government should . . .
   C. The surplus of wheat . . .
   D. Last year, exports of wheat . . .

Other exercises may call for identification of inferences supported or contradicted by data in a table, a graph, or summarized in a paragraph:

1. The data given in the graph could be used to argue that
   A. industrial profits are too high.
   B. price controls should have been retained longer after the close of World War II.
   C. the supply of consumer goods has decreased in recent years, hence, prices are high.
   D. a depression is inevitable.
2. Judging from the figures reported in the article, it is correct to conclude that
   A. consumers are buying more goods, but at lower prices.
   B. consumers are buying less goods.
   C. manufacturers intend to increase their sales efforts.
   D. one-fourth of the manufacturers are decreasing output although customers are plentiful.

An exercise may evaluate the ability to recognize the need for additional evidence or of the effect of new evidence in accepting or rejecting some conclusion stated in the quotation:

1. One could more justifiably accept the conclusion that slum children are less intelligent than children from more privileged environments if we knew that the investigator had
   A. tested larger samples of both types of children.
   B. used a test that is fair for both types of children.
   C. done both of the above.
   D. done neither of the above.
2. We could most justifiably accept the author's conclusion that the Taft-Hartley Act has served to "enslave" workers if we knew that
   A. labor leaders in general oppose this Act.
   B. opinions of plant managers were obtained.
   C. the status of workers had actually declined since its adoption.
   D. a survey was made of the opinions of a representative sample of workers.

Frequently, an exercise should call for the identification of the consequences or effects of given courses of action. For example:

1. If the sales tax advocated by the author of the above quotation were to be adopted, which of the following would be likely to happen?
   A. Less money would be collected by means of the income tax.
   B. People with large incomes would suffer more than people with small incomes.
   C. People with small incomes would suffer more than people with large incomes.
   D. The burden of the tax would be proportionate to income.

Conclusion

A variety of types of objective exercises have been discussed and illustrated and numerous suggestions have been made concerning the art of writing them. An effort also has been made to emphasize relationships between social science instruction and evaluation. The question may be raised as to how we know that such exercises measure more than recall information. It is evident from study of the exercises that knowledge is an important factor in their solution. It seems to me that it is also legitimate to infer that many of the intellectual skills functioning in critical thinking are also evaluated, assuming that social science instruction has been of the character described.

REFERENCES

1. Bishop, Hillman M., and Hendel, Samuel. *Basic Issues of American Democracy*. New York: Appleton-Century-Crofts, 1956. 484 p.
2. Bloom, Benjamin S., and others. *Taxonomy of Educational Objectives*. New York: Longmans, Green and Co., 1954. 192 p.
3. Chausow, Hymen M. *The Organization of Learning Experiences to Achieve More Effectively the Objective of Critical Thinking in the General Social Science Course at the Junior College Level.* Doctor's thesis. Chicago: University of Chicago, 1955. 138 p.
4. Dressel, Paul L., and Mayhew, Lewis B. General Education: *Explorations in Evaluation*. Washington, D. C.: American Council on Education, 1954. 325 p.
5. Hunt, Maurice P., and Metcalf, Lawrence E. *Teaching High School Social Studies*. New York: Harper and Brothers, 1955. 471 p.
6. Johnson, Earl S. *Theory and Practice of the Social Studies*. New York: The Macmillan Co., 1956. 476 p.
7. Lee, Raymond L., Burkhart, James A., and Shaw, Van B. *Contemporary Social Issues*. New York: Thomas Y. Crowell Co., 1956. 864 p.
8. Levi, Albert W. *General Education in the Social Studies*. Washington: American Council on Education, 1948. 336 p.
9. Sims, Verner M. "The Essay Examination is a Projective Technique," *Educational and Psychological Measurement* 8:15-31, Spring, 1948.
10. Stalnaker, John M. "The Essay Type of Examination," Chapter 13 of Lindquist, E. F., and others. *Educational Measurement*. Washington, D. C.: American Council on Education, 1951. p. 495-530.
11. Weinberg, Meyer, and Shabat, Oscar E. *Society and Man*. Englewood Cliffs, N. J.: Prentice-Hall, 1956. 782 p.

DISCUSSION

Participants: Eugene D. Carstater, Max Martin Kostick, Leo Nedelsky, Arthur E. Traxler, J. Wayne Wrightstone.

CHAIRMAN WRIGHTSTONE: We have a few minutes for questions, and although I am basically a friendly person, I have been advised that I shall have to prompt you if you do not speak loudly enough so that you can be heard by the stenotypist. I shall carry forward this assignment, and prompt you if you do not speak loudly enough.

DR. KOSTICK: This is for Dr. Nedelsky.

You suggested that an erroneous choice that is too sweeping is not good in so far as it penalizes a particular group of students. We might say the group that is penalized subscribes to sweeping statements while the group that is too conservative would have an advantage on such an item. Now another example would be a choice that would not be broad enough or sweeping enough. This type of item would penalize the group that thinks too broadly.

My question is, how do you feel about using a five choice item in which there is one best choice; the four other choices penalize four groups that are biased in different directions?

DR. NEDELSKY: I definitely agree that would be a proper solution. You can do it if you are ingenious enough to have it in one item. But it is perfectly all right to penalize one group in one item, and another one in another.

DR. CARSTATER: I was struck by the aptness of the story of Dr. Nedelsky's treatment of the archbishop's question. On being pierced, each of the speakers gave forth two substances. These illustrate the inter-disciplinary approach that has prevailed in these treatments. Diederich was to talk on the humanities, and a large part of his statement had to do with social sciences. I couldn't help observing that Nedelsky was getting into semantics, which seems to me to be a part of the humanities. And a large part of Engelhart's presentation had to do with the scientific aspect of the social studies. I think this is quite desirable, that there should be this interplay.

One other thing about the aptness of the illustration, however; the bishop asked, "Why?" And I am a little bit concerned with the possibility that we are now using the term "exercise" without too much look at what are we using the exercises to do. It seems to me that the appropriateness of many of the canons, or pseudocanons, with regard to how to write exercises or how to write test items, depends on what you intend to do with the results. The "how" depends on "what," and both depend on "why?"

CHAIRMAN WRIGHTSTONE: Does any member of the panel wish to respond to this statement?

DR. NEDELSKY: What he said was true.

CHAIRMAN WRIGHTSTONE: We accept the truth of this statement.

I want to express my own thanks to the panel members who have presented an excellent review of exercise writing in humanities, in science, and in the social studies. When you see these papers printed in the proceedings, I am sure that you will be able to study them much more carefully than is possible in an oral presentation. I have had the opportunity to see them, at least in mimeograph form, before the meeting. I want to thank each of the speakers personally, and on behalf of the group for his presentation.

LUNCHEON ADDRESS

69

Prediction of Educational and Vocational Success Through Interest Measurement

EDWARD K. STRONG, JR.

What is success? What are interests? How are they related?

In the days when tests were just beginning to be used, Edward Thorndike preached that "if a thing exists, it can be measured." Today when there are tests galore, he would probably say, be sure you know what you are trying to measure and then use only the test which is specifically designed to measure that characteristic.

Your assignment calls upon me to discuss the use of a test which was not designed to measure success. Consequently the conclusion is what should be expected—namely, interest tests do not correlate to any practical degree with measures of success. But I am convinced that interests have a real relationship with satisfaction.

Having answered the assignment given me I might be justified if I sat down, but I am afraid your program committee would not feel that I had earned my luncheon. Instead I am going to utilize this opportunity to outline some of the things we know about success and interests and to hint at some of the things we do not know. We can't survey the whole situation for we don't know enough to comprehend the total situation. There are hundreds of pieces in our jigsaw puzzle—pieces mentioned in hundreds of articles and books—and there are undoubtedly many pieces that are missing. But maybe we can obtain some glimmering of the complete picture which will be disclosed some day when all the pieces have been assembled. Since my time is limited, only those pieces can be mentioned which pertain to success, interests and the relationship between the two. And the pieces that are mentioned must be described in a few words without reference to all the ifs, ands, and buts that ought to be recognized.

Success — Satisfaction

What is meant by success? Success is an evaluation by one or more judges as to how well some activity is performed in terms of a standard or criterion. There are thousands of activities. One may be a success in one activity and not in another. For each activity there are one or more recognized standards by which success is measured. These standards range from very definite objective measures to very vague subjective measures. Because several standards are recognized, an author, for example, may be judged successful in terms of income, judgment of critics, or Nobel Award. Similarly a freshman may be called a success if he has an A grade average or even if he has merely passed the course. Furthermore, men are rated as successful or not in terms of a given population; thus an athlete may be rated the best high jumper in high school, in college, on the Pacific Coast, or in the world.

Men are acclaimed as successful and will continue to be so designated in everyday life. But the basis for such awards will vary from very objective to highly subjective standards and will differ appreciably among different observers. In any careful investigation there should be substituted for the vague term of success a definite statement as to what specific activity is being considered, how well it is performed in terms of a specific standard or criterion, who are the judges, and within what specific population the contestants belong.

There is another complication that muddies up our thinking. Every man is influenced to some degree by the evaluation of his behavior by others. But to a large degree his behavior is determined by his own evaluations. These two systems of evaluation lead, respectively, to success or satisfaction. What then is satisfaction? Satisfaction is a mode of consciousness accompanying activities by which one attains his own personal goals. Success is based on some standard recognized by others, satisfaction is one's own appreciation that he has reached his goal or is advancing toward his goal. Failure in these respects arouses dissatisfaction. All behavior whether contemplated or overt brings some degree of satisfaction-dissatisfaction. But only a few activities are successful in the eyes of others. There is some relationship between satisfaction and success, but how much is undetermined. If one's goal is to win the half mile then success is necessary for satisfaction. However a vast number of activities are carried on that must be judged as far from successful in the eyes of others but that nevertheless bring satisfaction. I, for example, like fishing, golf and gardening but I have never won a prize. I simply enjoy them, obtain satisfaction thereby. Brayfield and Crockett (2) report very little relationship between employee performance and employee satisfaction, as expressed by attitude surveys. Improved procedures may raise the low correlations somewhat. It is doubtful, however, if there is any high relationship, since only a modicum of success is necessary to hold most jobs. It must be recognized that many students and employees do not strive for great success in the classroom or on the job. They don't want to work that hard. But on some other activity they may expend a great deal of effort and energy. They want to enjoy life with their families in their homes, to have leisure time and

71

enjoy it in a multitude of ways. In other words, their interests help determine the amount of energy they are willing to put forth and how hard they are willing to work, and sometimes these interests find no outlet on the job. After all only a very few can be rated as real successes in the classroom or on the job but all may obtain satisfaction elsewhere.

How is success achieved? Success is achieved through the use of abilities and motivations. Many different capabilities are utilized in performing any specific activity. Motivation is complex, involving ambition, willingness to work, willingness to forego present pleasures for future gain, liking to engage in the necessary parts of the total activity, rewards of all sorts, etc. But if the competition is too keen or one dislikes what has to be done and finds too little satisfaction in it, one quits and looks for a situation where there is less competition or for other activities more to one's liking. All through life everyone is faced with the problem of which way to go, in which direction one can reach his goals with the least effort and with the most satisfaction.

What are the criteria of success? In the business world men receive preferment, increase in pay, and promotion at the hands of their supervisors. Many believe that supervisor's judgments are largely subjective, usually biased, sometimes unfair. Such opinions, or ratings, are criticised because they do not truly reflect amount of work performed since it is assumed by the worker that vocational success is a matter of production. It is quite possible, however, that many so called biased opinions reflect a better estimate of the man's usefulness than does his available production record. There is probably no more troublesome problem than the determination of adequate criteria of success in managerial positions and other complex jobs, such as selling. It is doubtful if there is a single case where experts have unanimously agreed upon the standards in terms of which success is to be appraised. In the case of many jobs in the office and shop the amount of work is determined, as on an assembly line. Willingness to do the monotonous task must be far more important than ability to do the work. As long as we think of success as amount of production we correlate test scores with some sort of production record and say we are measuring success.

Do we really know the criteria of scholastic success? The typical answer is mastery of knowledge. The instructor's job is to impart certain knowledge and to determine how well his students have acquired that knowledge. And it is the student's job to acquire this knowledge. Acquired knowledge is measured by grades. Then we claim that grades measure success and acclaim students as successes or failures on the basis of their grades without much concern as to the lack of agreement between so-called success in school and so-called success in later life.

How valid grades are is still a burning question. The validity must vary greatly among instructors and among educational institutions.

But the problem of what are the proper criteria of scholastic success goes far beyond such considerations. Here is involved the question: What is the purpose of education? Should the emphasis be upon acquired knowledge, or proficiency in its use? Knowing something does not guarantee it will be used. Or should the emphasis be upon appreciation, enjoyment throughout life? It is often said that the purpose of education is to teach students to think. Do instructors know whether students have increased their capacity to think? Does anyone know how to measure increase in capacity to think? Do we not want students to do more than to think; do we not want them to become excited, to bubble over with enthusiasm, and to imagine possibilities of using the material in unheard of ways? How much of our instruction is designed to propel the student out beyond the frontier of our own knowledge? And how in the world could an examination be constructed to test the students' newly aroused imagination? It has been said that the country needs more leaders. But we must not confine the meaning of this term to leaders along the accepted trails, we need those who will seek out and blaze new trails in science, philosophy and the like, creative leaders as well as executives.

When emphasis is put upon grades as the measure of success then almost inevitably emphasis is put upon the students who get good grades and little or no concern is shown about the students who are initially rejected for admission or later on flunked out. Should not the emphasis be directed toward preparing all men as far as possible for positions where they will be reasonably successful and satisfied? Here there is need to consider specific abilities, interests, motives, aspirations, goals—in fact, the whole area of values. When all this is appreciated it is apparent that there cannot be one standard or measure of success but many definitions of success appropriate to all the diverse people who carry on the myriad activities that our country must have performed.

What has been said so far might be summarized as follows. First, a man is successful according as society values his behavior; he is satisfied according as he himself values his own behavior. Second, each man seeks varying degrees of both success and satisfaction, constantly asking himself the questions: "Am I doing the right thing? Should I change? What should I do next?" The primary objective of education should be the orientation of each student rather than the acquisition of knowledge. But I do not want to give the impression that I am belittling the acquisition of knowledge—certain knowledge is essential for success and satisfaction in each activity. Third, we have no clean cut, unanimously agreed upon, criteria of scholastic or vocational success.

Interests

Let us now inquire as to the nature of interests and what possible relationships may exist between them and success.

Interests pertain to the liking-disliking of activities. Pleasantness, liking, belief, satisfaction are affective states of mind indicative of acceptance. Unpleasantness, disliking, disbelief, dissatisfaction are indicative of rejection. The terms pleasantness, liking and satisfaction are not clearly distinguished in every day usage. Our working definition is to use pleasantness-unpleasantness with reference to sensations, liking-disliking with reference to activities, belief-disbelief with reference to attitudes, and satisfaction-dissatisfaction with reference to feelings and emotions related to the approach to or arrival at goals. Interest is an aspect of consciousness similar to satisfaction but not as thoroughgoing an affective state as satisfaction.

What interests are possessed can only be ascertained by noting the activities that are liked or disliked. To say a man has musical interest merely means that he likes more activities relating to music than activities relating to other groupings of activities.

To some degree interests are indicators of abilities, not as indicators of amount of abilities but of which abilities are preferred. Thus a boy who is poorer in linguistic activities than in mathematics will seek an environment where he can use mathematics rather than English. Similarly he seeks a college major and later on an occupation which involve more activities he likes than he dislikes.

How are interests measured? No one has so far derived a method of measuring amount of interest. I am very doubtful if there is such a thing as amount of musical, or mechanical, or camping-out interest, except in the sense that amount means the number of liked activities which are classified under some rubric. Instead, scores on interest inventories give the differences in interests between this and that.

Three methods of measuring such differences in interests may be mentioned. Kuder has grouped interests into certain classes or rubrics, such as scientific, musical and computational. The testee must select which one among three rubrics in a triad he prefers. If he prefers one rubric to all the rest he can obtain a very high score but if his interests are more widespread he must necessarily obtain lower scores on each of his preferred rubrics. The Kuder does not then measure absolute amount of interest in each rubric but relative amount among rubrics.

A second method of measuring interests is that used with the occupational interest scores on my inventory. Here the interests of the average member of an occupation are contrasted with the interests of the average man.

The third method of measuring interests is employed with the interest maturity, masculinity-femininity and occupational level scales where the interests of men are differentiated from the interests of boys, similarly the interests of men are differentiated from the interests of women. Scales of this type have been constructed in which the interests of the best students are contrasted with the interests of the poorest students and similarly the interests of superior employees with the interests of inferior employees.

Measures of success can be appropriately compared with scores of superior-inferior interest scales providing the two purport to measure related abilities and interests. But measures of success cannot be appropriately compared with scores on the Kuder or my occupational interest scales since these do not purport to express differences in interests between superior and inferior persons.

It is a mistake to assume that scores on occupational scales measure amount of interest. Such scores indicate extent of agreement of the man's interests with the interests of the occupational group in contrast to the interests of men in general. Figuratively, the scores indicate direction to go. The higher the score the greater is the certainty that the individual should go in that direction. Direction to go is well portrayed by our Interest Global Chart where lawyers are located at the north pole, carpenters at the south pole, ministers far to the east and presidents of a manufacturing concern far to the west. Interest scores indicate direction to go, toward lawyer not carpenter, toward minister not president. Interest tests are not precise instruments like a compass. They can only indicate general direction. It is a serious matter if a man plans to become an engineer when his interests are those of a minister or social science teacher, for here the directions are largely reversed. It seems incredible that a man should so misunderstand himself. But I have myself counselled people who were intent on careers not in harmony with their school records, their interests or at times plain common sense.

Use of Interests Tests

Let us repeat that no relationship can ever be determined between interests and success because success is a mere abstraction. All that can be done is to correlate interest scores with measures of some criterion in terms of which an activity is judged. Consider first the relationship between interest scores and grades since grades are usually accepted as measures of scholastic success.

Ability and aptitude tests are useful in predicting what men can do, or better what they cannot do, but they are not very useful in predicting what activities a man will select. I still remember the college girl who majored in physics with a straight A record who refused every inducement to continue in graduate school. No, she wanted to marry and have a family. Interest tests function in reverse manner. They do not predict measures of intelligence, grades and achievement but they do predict the direction men will go.

There are some exceptions to this sweeping, over-all summary. For example, significant coefficients in the thirties usually result when scores on the psychologist interest scale are correlated with intelligence test scores and with scholastic grades. Such correlations are proof that there are interest items which so function. It is possible that if enough such items were utilized a scale might be developed which would correlate substantially with such measures. This is all the more possible if the test scores expressed interest in some restricted area of knowledge and were correlated with scholarship in courses similarly restricted to that area.

It is possible that there is actually a higher relationship between interest and achievement scores than so far reported. One reason for such a statement is that in many cases appropriate interest scores are greatly restricted in range yielding low correlations. There are, for example, many students with high engineer interest scores who have never studied engineering but there are few students with low engineer interest scores who have taken more than one engineering course. We might expect higher correlations among students in required courses where some students do not possess interests appropriate to the course. One such investigation concerned students who were ordered by the army to take a personnel-psychology course which most of them did not want (4). Here the data showed that interests correlated as highly with course grades as did intelligence test scores.

About the same relationship holds between interests and occupational success as between interests and academic success. Most of the research has been concerned with foremen, salesmen, production and sales managers, and research engineers. Correlations between some criterion of success and occupational interest scores have been for the most part negligible. There is one striking exception. Scores on the life insurance, realtor, and sales manager scales have correlated significantly with sales production and negatively with turnover among many groups of salesmen. It is quite possible that this relationship would hold for other salesmen if there were scales which more truly represented their sales activities. The validity of these salesmen scales may possibly be explained on the premise that salesmen must be really interested in their work if they are to succeed. It is possible that other occupational scales might correlate significantly with measures of success if adequate measures of success were available.

Occupational interest scales were not designed to predict success. If it is desired to predict success then interest scales should be developed contrasting the interests of the best employees with those of the poorest men. Research with small samples suggest the feasibility of such an approach. But actually, it is difficult to obtain large samples of inferior men. Although secondary school administrators complain about unsatisfactory teachers, a serious effort by Gilbert Wrenn and myself to obtain the names of such teachers made clear that high school principals in California would or could not supply more than 62 names. It is to be hoped that the Life Insurance Agency Management Association may attempt such a scale. They, if anyone, can secure the names of many unsuccessful salesmen.

A more feasible program would be to contrast the upper twenty-five per cent of an occupational group with men in general on the assumption that the average man is no more fitted to be a success in that occupation than those who have been found wanting after a tryout. This would be equivalent to occupational scales based not on the average member of an occupation but on the superior members of the occupation. Such an innovation might be more useful in selection but it is doubtful if it is desirable for guidance.

Turn now from the relationship between interests and success to the relationship between interests and orientation of students and employees.

There is no question but that occupational interest scales differentiate occupational groups. But do occupational interest scales appropriately differentiate students and what validity do scores on these scales possess in the sense of predicting the occupations in which students will eventually engage?

Although my test was not designed to differentiate between college curricula or majors, Berdie (1) reports that interest scores predict such choices better than either aptitude or achievement tests. This must be because men choose curricula which they deem appropriate to their occupational choices, or vice versa.

Graduate medical and business students at Stanford are very clearly differentiated by physician interest scores. Seventy-three per cent of medical students obtained an A rating in contrast to one per cent of business students and only one per cent of medical students had a C rating in contrast to 83 per cent of business students. It is doubtful if these two groups could have been differentiated to any practical degree on the basis of aptitude test scores and undergraduate scholarship records.

Occupational interest scores not only differentiate occupational groups but they predict the occupations men will be engaged in later on. The data from an 18-year follow-up of Stanford students show that if a student had an A rating on any one of 16 scales the chances were 78 to 22 he would actually be engaged in that specific occupation 18 years later and that if he had a C rating the chances were 83 to 17 he would not be so employed (5). The expectancies are really better than the above figures indicate since all men employed in a closely related occupation were counted as misses. Thus all men with an A rating on the accountant scale who were engaged in office, credit or purchasing work or even as CPA's were not considered to be accountants. So far no satisfactory method has been discovered whereby credit can be given to those who entered occupations closely related to the occupation in which they were tested. This is unfortunate since high scores on an occupational scale should always be interpreted as favorable to related occupations.

Occupational interest scales have been devised primarily to differentiate between gross occupational groupings, as, for example, chemists, engineers, physicists and mathematicians. Recently it has been demonstrated that sub-groups of engineers, of psychologists and of medical men can be similarly differentiated on the basis of interests. Several hundred medical students have been scored on the four medical specialty scales and a ten-year follow-up is planned in order to determine how well such fine discriminations predict the specialties in which medical students eventually engage (6).

There are many orientation problems facing young people. Interests must play a role in many of them. Which students should enter the unskilled trades, which the general run of business activities, and which pursue graduate work are not questions pertaining merely to ability. Many competent high school graduates do not go to college because they don't want to; many college men enter occupations of lower status than their IQ would predict.

There are certainly many men who really enjoy their work and they are to be found in possibly every occupation. I once employed a gardener with an IQ of not over 80, whose eyes would light up with pleasure when I asked him to spade a tough area or to dig up an old stump. We don't know, however, whether all men are so constituted that they can enjoy appropriate work; also whether the number and variety of jobs are proportional to the number of men with corresponding occupational interests.

We don't know how many adults read books for pleasure who never went to high school; we don't know how much high school and college has increased reading for pleasure, or even decreased such enjoyment. And we don't know whether those who read for pleasure differ in abilities or interests or both from those who don't read books. Furthermore, we don't know to what extent English can be taught so that people will read for pleasure and whether such instruction can influence all students or only a few, and if but a few, how they differ from the rest.

Several law and medical faculties are worried about certain alumni who are interested in making money or are engaged in non-ethical practices. They want to know how such future alumni can be identified and eliminated. It is doubtful if intelligence tests and grades would differentiate such students to any worthwhile degree.

These examples are only samples of many problems regarding the differentiation of young people, of aiding all young people to find a career in which they can be reasonably successful and satisfied.

It is generally assumed that people with high IQ can succeed in any career; consequently it is of minor importance what career they espouse. This is largely true, such people *can* so perform; but should they? A considerable number of men in their forties and fifties, have sought my counsel. They had come to realize that their work was work, dreary work, and they were tired of it; and had discovered that some of their friends found their work interesting, exciting. They reminded me of driving an auto with the hand brake pulled back, when one is wasting gasoline and wearing out the engine and brakes. These men were tired, bewildered and one was on the verge of a serious mental breakdown. They all scored low on interest scales appropriate to their work. The objective of my research during 33 years has been not merely to direct young people into work they will enjoy but to prevent older people from wearing themselves out doing work they don't like and being forced to continue therein because in most cases it is now too late for them to enter a new career. The present day interest in retirement and the appalling number of mental cases in our hospitals appear to be to some degree the resultant of people forced to work at that which is foreign to their interests.

May we repeat in summary that success is obtained by skill, knowledge and drive. One's interests are not particularly related to how much skill or knowledge one is capable of acquiring but they are closely related to the drive that stimulates one to make use of all of his capacities. It is in the orientation of students and employees toward their most productive activity that interests play a most important role.

REFERENCES

1. **Berdie**, R. F. Aptitude, achievement, interest, and personality tests: a longitudinal comparison. Journal of Applied Psychology, 1955, 39, 103-14.
2. **Brayfield**, A. H. **and Crockett**, W. H. Employee attitudes and employee performance. Psychological Bulletin, 1955, 52, 396-424.
3. **Kelly**, E. L. **and Fiske**, D. W. The Prediction of Performance in Clinical Psychology. Ann Arbor. University of Michigan Press, 1951.
4. **Strong**, E. K., **Jr.** Personnel-psychologists at Stanford University. Psychological Bulletin, 1944, 41, 474-78.
5. **Strong**, E. K., **Jr.** Vocational Interests 18 Years After College. Minneapolis, University of Minnesota Press, 1955.
6. **Strong**, E. K., **Jr. and Tucker**, A. C. The use of vocational interest scales in planning a medical career. Psychological Monographs, 1952, 66, No. 9.

SESSION III

Test Users' Problems as Guides to Better Measurement

Remarks of the Chairman

ARTHUR E. TRAXLER

Ever since the beginning of the use of standardized tests, specialists in the field of measurement have had a good deal of free advice for test users. They have admonished users to choose tests which are designed to measure the major educational objectives of the schools, to administer these tests with faithful attention to the standard directions, to employ scoring procedures that will assure accuracy in the scores, and to interpret the results in the light of statistical data on the reliability and validity of the measures. Much has been said and written concerning the "education of schools in the meaning and use of test results."

At the same time, test specialists have, in recent years, become increasingly aware of the need for communication between producers and consumers of standardized tests to be a mutually active process. They have learned that teachers, counselors, and school administrators have much to contribute to test production. Many of them have, in fact, come to recognize that school personnel must take a leading place in determining the objectives, scope, and content of the tests which are used with students. So the current trend is toward the development of new tests as a team effort by administrators, teachers, counselors, and test specialists working in close cooperation.

This afternoon's session is in line with the modern desire of test specialists to learn about measurement needs from test consumers. First, we shall hear from an administrative officer of a large city school system who has himself made numerous contributions to measurement and research. Later, a national authority in the guidance field will present a counselor's view of measurement needs. Each paper will be discussed by a leader in the field of educational measurement. While the papers are being presented, members of the audience are invited to note any questions or comments for use in general discussion near the end of the session.

The School Administrator's Problems for Testers

PAUL T. RANKIN

Tests and testing are of continuing concern to school administrators. This afternoon I propose first to review the uses of tests from the point of view of the school administrator, and second to express some of the problems in the field of testing as seen by the school administrator, in the hope that specialists in testing may be able to help meet these problems more adequately.

The School Administrator's View as to the Uses of Tests.

A general function of school administration is to provide the facilities and means whereby instruction can be carried on. Tests constitute one such means. Probably the primary use of tests, from the point of view of the school administrator, is *to help teachers know and understand children and their needs better*. In the last twenty-five years, there has been a growing emphasis on the importance of having teachers not only know what there is to know about child growth and development in general, but also have control of procedures by which they can become better acquainted with each child's capacity, interests, and accomplishments in every major field with which the school is concerned. In the Detroit school system last year, with a total membership of 275,000, about 96,000 intelligence tests and 1,300,000 printed achievement tests of one sort or another were used. Of the total number of such achievement tests, approximately 1,240,000, or 96 per cent, making up what we call the optional testing program, were made available to teachers at their request for use in their classrooms at the time and in the way they wished. No reports of the results of these tests were made centrally. The tests were provided solely to enable teachers to compile information about each pupil's status in relation to some instructional goal. The other phase of achievement testing is the guidance and counseling testing program. This consists of the Iowa Multi-Level Test of Basic Skills and a locally developed inventory of experiences and interests which are administered in grades Low 4, Low 6, and Low 8.

A second use of tests, from the point of view of the school administrator, is to enable teachers *to evaluate better the accomplishment of pupils in particular courses*. This has reference especially to the secondary level, where promotion takes place usually by subject rather than by entire grade. In most cities, I believe that tests are rarely if ever used as the sole criterion for satisfactory completion of secondary school courses, but standardized tests can be and are being used to afford additional information to teachers as to the progress made by students in particular courses. This use of tests is even more important when instruction is given by special means such as correspondence or television.

A third important use of tests is *to help in the counseling and guidance of children, and particularly in classifying them for instructional purposes.* Tests both of capacity and achievement are needed for this purpose. Let me refer again to our Detroit experience. The results of intelligence tests and the achievement test battery provide information which is helpful to counselors in the junior and senior high schools in guiding students into programs best suited for them. Such use is a major reason certain tests are given just before the end of the child's experience in the elementary and the junior high school.

Test results are used extensively in the classification of pupils—both for assignment to particular courses and classes and for ability grouping within a particular class. For example, the decision as to whether a ninth-grade student takes college preparatory algebra, general mathematics, or arithmetic review depends very largely on the results of the tests that have been administered in the eighth grade. Likewise, the sub-groups in a second-grade class in reading or an eighth-grade class in civics or a twelfth-grade class in English may be based on test results.

A significant phase of the guidance use of tests is as a resource in conferences with parents. The school that has a good background of testing data for pupils is prepared to explain objectively to parents the reasons for proposing particular courses of action. In this connection, the Detroit achievement profiles of children at the fourth-, sixth-, and eighth-grade levels give considerable information.

A fourth use of tests, from the administrator's standpoint, is *to answer the question of how well the school or school system is doing in instruction generally, and to aid in the redirection of emphasis in the instructional program.* Members of boards of education and citizens' advisory committees continually ask how the test results in the local school system compare with those in other school systems. This is a very real question, and one that is exceedingly difficult to answer fully and accurately. Truly comparable data are difficult to secure. Pupil populations in different school systems vary widely, grade for grade, in general capacity, in home environment, in age, in time allotments, and indeed in every factor that affects academic success. Interpretation of differences in test results between two school systems calls for great care. Furthermore, when the comparisons are unfavorable to the local school system, there is a tendency to withhold this information from the general public. The difficulties in dealing with the question, however, do not make it any the less important. The superintendent of schools who is charged with the general overseeing of the entire school program needs information regularly as to the attainment at the various grade levels in comparison with attainment in comparable schools elsewhere.

Knowledge of the comparative results in various fields of instruction can aid the administrator materially in redirecting emphasis on an intelligent basis. For example, if the test results indicate that at the end of the elementary school period the school system is accomplishing good results in comparison with other school systems in the fields of reading and spelling but not in arithmetic or skills in social studies such as map reading, a shift of emphasis may be indicated—either by a change in the time allowance or in the selection of materials, or by some other means. The problem of reliable and truly representative norms is of great importance in this use of tests.

A fifth use of tests is *to answer the question of how the results being secured now compare with the results secured five or twenty or fifty years ago.* This point is raised periodically by people in the community. Often the assumption is that children do not learn as much as they did X years ago. Whether or not this is so is a question which every school administrator would like to be able to answer satisfactorily. Practically, however, this is an exceedingly difficult question to answer, because test results of the past usually are not available. Or, if such information is available, the presumably comparable tests of today have changed content and may not be truly comparable. Or emphases in instruction have changed; no tests were given at an earlier period to measure attainment of objectives considered important now, and consequently comparisons are incomplete. Thus, it seems that test makers would do a great service to school administrators by providing the results of then-and-now tests, where these are determined in school systems able to carry out such studies. But care must be exercised to insure that the pupil populations then and now have similar characteristics in terms of general social and intellectual level. Comparisons between substantially changed pupil populations may render a disservice to the profession generally.

A sixth use of tests in school systems is *to help in the selection and promotion of teachers and other school employees.* This is a relatively recent but rapidly growing development in many school systems, including Detroit. Tests have become an important element in personnel administration. In Detroit we now use tests as a part of the process in the initial selection of teachers, and in promotions to department head, counselor, and assistant principal. Tests are used in the selection and promotion of a wide variety of non-teaching employees, including stenographer, clerk, bookkeeper, janitor, cleaner, and lunchroom worker. There is need for more valid and predictive instruments in all theee areas.

Problems in the Testing Field

Perhaps the central problem in the field of testing, from the administrator's viewpoint, is the reluctance of many teachers to use tests as fully as their value would appear to warrant. What are the reasons that many teachers, and some administrators, have for not using tests as extensively and as fruitfully as seems desirable?

A first group of problems centers around the *time, effort, money, and inconvenience involved in using tests.* Teachers and administrators alike want tests which will result in maximum information with the minimum expenditure of time and effort and with the least inconvenience to school staff and school program. In this connection, one often-expressed need is for tests that can be handled in one or more normal class periods. Having an uninterrupted test that requires over 40 minutes imposes a problem at the junior and senior high school level because the traditional length of the class period is only 45 minutes. Tests are needed which can be administered, in their entirety or in parts, in time units of no more than 40 minutes.

Another point made by teachers is that tests should be provided in formats which can be used comfortably on the smallest working surface usually provided, since many classrooms are equipped with tablet-arm chairs rather than the older type of desks with large surfaces. Teachers feel that test booklets are more usable in such situations when pages can be turned and folded under, thus taking up a minimum of work space. With booklets that have many pages, some type of ring or spiral binding is desirable.

Teachers and administrators alike prefer editions with separate answer sheets from the standpoint of economy, both in initial cost of the tests and answer sheets and in time and cost of scoring.

Adaptability to machine scoring is another point to be considered in insuring more extensive use in the schools. A year ago the Detroit schools adopted as the major test in their guidance and counseling test program the Iowa Multi-Level Test of Basic Skills because, among other reasons, mechanical scoring was available at reasonable cost and with fast service. Two times last year, answer sheets to the number of nearly thirty thousand were sent from Detroit, scored, and returned with results within three or four weeks. The fact that the actual scoring of these very long tests did not need to be done by the individual teachers figured greatly in the teachers' willingness to use the tests and the results as fully as possible.

The use of the same test booklets a number of times calls for durable construction and materials to resist to a maximum degree the normal wear and tear of ten or fifteen administrations.

Another suggestion often made by teachers and principals is that the author should offer extended suggestions for interpretation and use of the test results. Too often, from the beginning of the testing movement in education, there has been a tendency to give a test, look at the results, and then file them with little if any consequent action. Today the situation is much improved in this regard, but the test maker still has an important responsibility to report to prospective test users what other test users have found helpful in utilizing test results to greatest advantage. Such suggestions may be included in the test manual or in special supplementary bulletins devoted exclusively to ways of following up and using the test results.

A second major problem that administrators see with regard to tests is *the need for units that are comparable from field to field and from grade to grade*. This problem is not a new one to the specialists in testing. You have wrestled with it for many years. You report test results in grade levels, in various kinds of so-called standard scores, in percentiles, and in other ways. It is obvious, of course, that whatever unit is used has meaning only in terms of the adequacy of the sample of the population on which the norms are based. The question to which I wish to address attention here, however, is the selection of the unit itself. I think school administrators would be glad to have all achievement test results in the elementary and secondary schools expressed in the same units, so that comparisons may be made from test to test for particular children at one time, and for successive administrations from grade to grade. I urge continued attention to this problem of finding the unit that meets best the varied requirements of a good unit in which to express test results.

A third problem that concerns school administrators is *the need for norms that are based on adequate and representative data for various types of population*. For example, two different reading tests are administered to High 7 pupils. Even though the tests appear, at least superficially, to measure the same group of reading abilities in substantially the same way, the results on one test may indicate that the pupils are a half-year below the norm, while on the other test they may indicate that they are a half-year above the norm. Situations of this type give rise to grave doubts in the minds of teachers and school administrators. This in turn hinders the wider use of tests by school personnel and the fuller development of the testing movement in school systems.

There are problems also in the adequacy of norms in the case of some tests. Several years ago in Detroit we tested seven thousand students in a high school subject with the only test commercially available that seemed suited to our need. But the norms offered by the author were based on a population of fewer than four hundred students.

In this connection I should note also that school administrators welcome differential norms for different types of school populations. Again, let me refer to our experience in Detroit. Ours is a highly industrial city, with a changing population. We would like to have norms available for the tests we use which would represent average achievement in cities similar to Detroit, and possibly even for type areas within large cities, rather than averages for states with predominantly rural populations or cities with characteristics markedly different from those of Detroit. To repeat, we need norms for every test that are truly representative of various types of populations.

The fourth problem to which I want to direct attention is *the need for achievement tests for every major course—especially at the secondary school level.* This year, for example, Detroit is participating in an experimental project, sponsored by the Fund for the Advancement of Education, designed to evaluate the teaching of large classes by means of television. We are conducting the experiment in six subjects at different grade levels. One of these is American literature in the senior high school. Our instructional research staff checked carefully and were unable to find a standard test that was judged to be satisfactory for our purposes in comparing results in classes in American literature using the television course and those conducted in the usual manner. There is a real need for tests in every major field of instruction at each of the different levels of maturity.

In Detroit we have been making extensive use of tests in reading, arithmetic, and spelling for forty years. From time to time we have used different tests in the social studies, in English, and in some phases of home economics and industrial arts. It is worthy of note, however, that we have only rarely used tests in art, music, health education, physical education, or science, although these fields constitute a large part of the curriculum in grades 1-12. In recent years we have used relatively few standard tests in English expression and in the social studies, although these fields of instruction take a substantial share of the instructional time. Of course there are varied reasons for this, but one is that we, at least, have not been able to locate tests that seem to

88

serve our purposes. There is need for more, and more adequate, tests in these other fields.

The fifth problem I want to mention is *the need for tests—or groups of tests—which will measure more comprehensively the achievement of pupils toward all the outcomes teachers seek through the medium of the subject they teach.* Let me refer again to the situation in the teaching and testing of American literature. The teacher of American literature has a number of objectives in mind. These include acquaintance with writers and writings, creation of the desire to read American literature, and development of taste and discrimination in literature. A truly adequate test of pupil growth in American literature would evaluate not only the knowledge pupils have acquired about American writers and their writings, but also the degree to which they have developed taste and discrimination in American literature. Ideally, such a test would provide some indication of the likelihood of the students tested to continue reading American literature after the course is completed. This is what I mean by a test—or group of tests—that measure more comprehensively the attainment of pupils toward the schools' total group of objectives.

One final problem I want to present to you. This is *the relationship of standard tests to the marks that teachers assign pupils for their work in the school subjects.* Pupil marks, and report cards, are the subjects of continuing debate practically everywhere in this country. Parents generally are not too well satisfied with the marking systems that are currently in use. Neither are teachers. The pupil's mark may represent his level of ability on some absolute scale, or his ability in relation to the grade as a whole, or his ability in relation to the particular class, or his achievement in relation to his capacity or effort—or some combination of all of these criteria.

Is it possible for the testers to come up with practicable ways in which objective test scores may be used to help make pupil marks more meaningful? Can the test scores be expressed in a form that will give parents a better idea of pupil achievement? Is there perhaps some modern adaptation of the old accomplishment quotient which may be used to supplement the absolute level of achievement revealed by the standard test and thus give some estimate of the student's accomplishment in relation to his capacity to achieve—either his general capacity or his specific capacity in a particular area? Certainly this problem is a real one in all school systems. I state it here in the hope that the specialists in educational testing may be able to offer some help in this area to school administrators.

You will note that I have said nothing here about validity, reliability, or equivalence of forms. School administrators rarely talk specifically about these technical characteristics of good tests. My omission of reference to them, however, does not betoken any lack of regard for their importance. Certainly the good test must be valid, and must be reliable; alternate forms must be equivalent. I have rather attempted to give expression to ideas that I believe are uppermost in the minds of school administrators as they consider testing in relation to the schools in which they work.

Discussion of the School Administrator's Problems

ROGER T. LENNON

Dr. Rankin, we of the test-making fraternity are indebted to you for helping us to see, through the eyes of the school administrator, the functions which our instruments serve in the schools, and for your thoughtful recommendations as to ways in which we may make these instruments more useful. The problems which you have outlined for our consideration here this afternoon are, I know, by no means peculiar to Detroit, but reflect experiences common to school administrators across the country. I trust that I shall be able to reflect the reactions of the test producers as faithfully as you have reported the feelings of school administrators.

I am immediately impressed by the fact that the problems to which Dr. Rankin invites our attention as test-makers are not new problems, but rather ones that have been persistent matters of concern for at least as many years as I have been associated with the testing field—now more than twenty. Can it be, I have asked myself, that the test-makers have been so unresponsive to the needs of test users that no substantial headway has been made in dealing with these matters? How valid are the reasons which we have traditionally advanced for being unable to comply more fully with such requests as Dr. Rankin makes of us? What are the prospects that we shall, in the near or more remote future, be able to give Dr. Rankin and his fellow school administrators some of the help he has here asked for?

Time, Effort, Money

Dr. Rankin first called for improvement with respect to what we may term the engineering aspects of tests—physical characteristics, time requirements, simplicity of handling, machine scorability, and the like. Test makers would claim, I am sure, that they have been acutely conscious of such considerations; that they have made notable strides during the past twenty years in the directions Dr. Rankin deems desirable; and that, if some tests now being produced fall short of what the consumer would consider ideal with respect to these characteristics, it is not because publishers underestimate their importance, but because of the many other factors which must also play a part in test design.

Let me be specific. There has been a decided effort to fit tests, or work units of them, to the 40-minute limit Dr. Rankin proposes. But considerations of reliability and validity sometimes argue for a longer test, whereupon a practical decision needs to be made as to the best compromise between administrative convenience and adequate measurement. What is the test-maker to do when he discovers that to produce an acceptably reliable measure in a given area requires a test of some 50 or 55 minutes duration? The conscientious test-maker will not sacrifice reliability for the sake of staying within a 40-minute time limit, nor does he want to extend the test from 50 to 70 or 80 minutes merely for the sake of filling two class periods. The time required for a test is, after all, largely a function of the complexity and scope of the ability or abilities being measured, and is not entirely to be set at the discretion of the test maker.

As to format, it is not happenstance that so many tests are printed as 8½ x 11 booklets. Considerations of effective utilization of page area, attractiveness and clarity of page layout, and size of sheet which can be handled economically on available presses, favor the 8½ x 11 page size, or one fairly close to it. When a booklet of this page size is used with an answer sheet, whether of the IBM or Iowa variety, the total area covered by booklet and answer sheet necessarily exceeds that of the tablet-arm chair of which Dr. Rankin speaks. The cumbersomeness of materials appears to be an inconvenience that most students, at least at the high school and college level, quickly learn to cope with.

Use of Separate Answer Sheets

The matter of adaptability of tests to machine scoring, through use of separate answer sheets, continues to be deserving of the most careful scrutiny, from the standpoints of economy, administrative convenience, implications for validity and effective utilization of results, *and* impact on the economics of test publishing.

I should comment first that our experience is not in accord with Dr. Rankin's report that "teachers and administrators alike prefer tests which may be used with separate answer sheets." In the case of those tests which World Book Company makes available in both versions—with and without separate answer sheets—there is no single instance in which the separate-answer-sheet edition is the more popular. In fact, there is fairly general resistance on the part of teachers and school administrators to the use of separate answer sheets at the lower grade levels (up to the fourth or fifth grade) because of the materials-handling problems that these present to very young children. At the upper grade levels we find, too, some lack of sympathy with use of separate answer sheets on the part of teachers who prefer to have the actual record of the child's work before them rather than just an answer sheet, which is ill-suited to instructional or diagnostic application.

The matters of relative economy, accuracy, and speed of hand vs. machine-scoring, and the effect on test-construction and utilization practices of use of separate answer sheets have been ably discussed elsewhere; and I shall not comment on these issues here other than to observe that the case for machine-scorable, separate-answer-sheet tests is not in every instance as overpowering as Dr. Rankin seems to be saying. I must not, however, neglect to call your attention to certain economic problems implicit in the growing use of separate answer sheets. It is obvious that use of a separate answer sheet, costing a fraction of what a test booklet costs, coupled with repeated use of a test booklet, means a considerably smaller return to test publisher and author per student tested. You can readily imagine the dismay with which I listened to Dr. Rankin's plea for test materials that will withstand the wear and tear of "ten or fifteen administrations." If this trend is to continue, then we must ask where the funds are to come from for production of tests of the quality that school people are coming to demand.

I suspect that few test users realize the magnitude of the costs involved in test development. The production of a major achievement battery, for example, calls for extended curricular research and textbook analysis before a single item is written; then for the writing and experimental tryout of perhaps as many as 15,000 items, and for nationwide standardization, not to mention the associated research undertakings, such as investigations of reliability, equivalence of forms, and the like—calling for an investment well into six figures. How many tests will have sufficient sales to sustain such developmental costs if the per-examinee return is of the order represented by the average price of an answer sheet or two, plus a tenth or fifteenth of the price of a booklet?

We understand the schoolman's desire for economy, but we feel that it is necessary to view the matter of expenditures for test materials in proper perspective. As nearly as our publishing industry figures reveal, the average expenditure per elementary and secondary school pupil for testing materials is in the neighborhood of 20 cents—certainly not more than 25 cents. Imagine! We spend $300, $400, even $600 a year to educate a child and then worry about saving a few cents on test materials. What other expenditure yields commensurate benefits to teacher, counselor, administrator and pupil? I say to you, not entirely facetiously, that what is good for the test-publishing industry is good for the test user. Exaggerated striving for economy in test materials can only dry up the sources of new and better instruments.

Comparable Units

The second problem raised by Dr. Rankin is *the need for units that are comparable from field to field and from grade to grade.* He urges attention to finding "the unit that meets best the varied requirements of a good unit in which to express test results." We can readily agree that the variety of ways in which test scores are interpreted is sometimes puzzling to the teacher, or even administrator; and we sympathize with the desire for simplification and uniformity. However, the appropriateness of a particular unit in which to express scores depends on (1) the nature of the ability or function being measured; and (2) the use to be made of the scores. For certain functions, expression of scores in terms of age or grade scales is natural and meaningful; for others, such units would have little if any value. Percentiles, standard scores of whatever variety, or any other type of unit of which I am aware, possess properties which render them suitable for certain types of tests, unsuited for others. The abilities we measure are so diverse, and the uses we make of our tests are so varied, that I regard the universal, all-purpose unit as a will-of-the-wisp, pursuit of which is foredoomed to failure.

Adequate and Representative Normative Data

Dr. Rankin called, entirely properly, for provision of normative data based on adequate and representative samples of various types of populations. I am sure that there is no test for which normative data adequate for all purposes are available, but manifestly progress has been made in this respect, and certainly there are now available tests and batteries of tests with quite respectable bodies of normative data.

Dr. Rankin cited the vexing problem of systematic differences between results of tests purporting to measure the same function, presumably attributable to the fact that their norm groups are not comparable. This is admittedly an unfortunate state of affairs. It is an unintended consequence of our independent, sometimes competitive, system of test development, and of the fact that tests are standardized at different times and on different populations. There are some grounds for hoping that this situation will be improved. First, I would cite improvements in norming technology; there is, for example, a growing body of information on community and school system characteristics related to test performance that will help us in the selection and more precise definition of norm groups. Secondly, and more specifically, the major test publishers are now engaged in a joint enterprise, under the aegis of the Committee on Tests of the American Textbook Publishers Institute, which looks to the development of an "anchor test," or series of tests, to be included in the standardization programs for all our respective tests, and to serve as a common defining instrument for their various norm groups.

Meanwhile, many of the publishers do provide data on the comparability of results of certain of their tests, as for example, School and College Ability Test vs. American Council Psychological, old and new editions of California Achievement Test, and Metropolitan Achievement Tests vs. Stanford Achievement Test. It is well to mention, however, that "comparability" or convertibility, is not an absolute property of test scores; scores which are found to be comparable, or equivalent, in one population will not necessarily be so in another—and hence the need, ideally, for local determinations of comparability. And I need not remind you that differences in norm groups are not the only sources of lack of comparability between tests.

Desirability of Many More Tests

Dr. Rankin calls for production of achievement tests for every major course, especially at the secondary level, and for tests or groups of tests "to measure more comprehensively the achievement of pupils toward all the outcomes teachers seek." The question of what tests will be developed and published is largely an economic one—almost entirely so as far as the commercial test publishers are concerned, and not negligibly so even for our "non-profit" associates. You may be sure that test publishers will be vying with one another to produce tests in any field in which the demand is of a magnitude sufficient to make the venture commercially worthwhile. Indeed, it is true of test publishing as a whole that it goes well beyond the publication of items that conservative business practice might suggest. I wonder how many educators realize that the very existence of the test-publishing industry depends upon the sales of not more than thirty tests or series of tests and that these tests, in effect, carry the hundreds of other test publications?

As rapidly as school people demonstrate that they will buy tests in given areas, those tests will get published. If, however, author and publisher must contemplate on the one hand substantial investments of time and money to develop tests to meet ever more exacting standards, and on the other hand rewards that are minimized through use of separate answer sheets, yielding a meager return per examinee, you must not be surprised if many desirable tests are left unauthored and unpublished.

I trust, Dr. Rankin, that my reactions to your comments have not appeared excessively negative and that I have not appeared to dismiss lightly any of your proposals. We publishers know that our continuing success depends on meeting satisfactorily the needs of you and your fellow educators, and believe me when I say that we listen with utmost attention to all your recommendations. It was Winston Churchill who remarked, "I have derived continued benefit from criticism at all periods of my life, and I do not remember any time when I was ever short of it." Dr. Rankin, I know that your remarks to us test makers have been intended in no critical sense; but I am certain they have been of benefit to us, and that the benefits will in time accrue to the school administrators whose test-using problems you presented so well.

The Guidance Director's Problems and Suggestions for the Test Specialist

EDWARD LANDY

I

Perhaps the best way to answer the question implicit in the title of this talk is first to state the problems which confront the guidance director as he goes about his job on a day-to-day or year-to-year basis. Then we can ask ourselves the question as to whether the solution of these problems can be helped through the use of tests. If the answer is "yes," as I believe it is in much of our work, then we have to ask ourselves in what ways present tests are inadequate and what can or ought to be done to reduce this inadequacy.

Let us take a look then at some of the common tasks or problems confronting the guidance director. (Obviously not every guidance director will be faced with all of these problems. Some of the problems may be the responsibility of other personnel in a particular school system.) We can begin at the beginning with a policy for school entrance age to kindergarten or grade one. Should all children be admitted on a rigid chronological age basis or should we try to take account of individual differences and have a flexible entrance age policy? Let us assume that, for a variety of reasons, we adopt a flexible school entrance age policy. We are still left with the question as to how we can determine whether a given child is ready to enter school. Peter may be ready but not Charlie. How can we tell?

Obviously we cannot tell by simply looking at the child or even by obtaining a fairly complete history on him. We need some measures which will enable us to discover strengths and weaknesses in the child's ability to learn, in his learning achievement to date, in his desire to learn and in his social maturity. In other words we need measuring instruments which will enable us to engage in a diagnostic appraisal of this child as a candidate for school entrance and thus be able to predict whether or not he is a good risk for admission.

Are there any measures available to do this job? At present I know of no single measure or test which will provide an answer to this problem. Nor do I know of any combination of measures the results of which can be compared objectively with a later criterion of success. We do have measures which can be administered by a clinical psychologist. These can be combined in a clinical judgment with other data to indicate strengths and weaknesses and to make a prediction as to potential success. But this is a very expensive process and as a practical matter is out of the question for most school systems. Can the test experts develop some reliable and valid instruments which do not need clinical training to administer and interpret and which will be useful in this problem?

Let us move along in the educational stream whose headwaters we have just taken a peep at. Here is Susan in Grade III who is not doing as well as the rest of her classmates. A reading test indicates that she is reading on a low second grade level and the teacher's judgment agrees with this estimate. Mother has become very anxious about Susan's learning and has, quite rightly, brought considerable pressure to bear upon the school to do something about the problem. Why is Susan not learning well? Here again we have a problem in diagnosis. Is it because there are gaps in previous learning which are now handicapping her and which can be corrected by some good remedial teaching? Fortunately in the field of reading there are tests which do provide us with substantial clues as to lacks in skills and knowledge. Unfortunately this is not true of other areas of learning, except possibly for some of the simpler skills in arithmetic.

Let us assume that the particular gaps in Susan's learning are discovered and remedial teaching is tried to overcome them. Let us further assume that Susan continues to have difficulty. Here we have a problem of attempting to assess Susan's capacity for learning. Here tests are useful if they can tell us that Susan is or is not capable of learning any better than she is. We also need to know this as specifically as possible. Is Susan's difficulty with learning to read the result of a language disability which interferes with "whole word" learning thus making it necessary to use a detailed and rigorous phonetic approach? Or is too much anxiety the disabling factor? We need to know also whether Susan is willing to learn or, to put it differently, has resistances which interfere with the desired learning.

This type of problem applies all through the grades. More and more in our schools the counselor is expected to help pupils who have learning difficulties. Very often we are not sure as to whether the learning difficulty is due to lack of motivation or resistance to learning or whether there are some specific difficulties in learning which act as blocks to further learning. Tutorial or remedial techniques would apply in one instance and psychological counseling in another or both may be needed. I would guess that currently there is a terrific waste of remedial or counseling time which at best is being used to find out that an alternative technique should be employed. Under ideal conditions at present this is being done by clinically trained psychologists but obviously this is very expensive and impossible for most school systems.

Again can instruments be developed which do not need the clinically trained person to administer and which can do the twin jobs of diagnosis and prediction? I hasten to add that even the clinically trained psychologist has his difficulties as well and would undoubtedly welcome more reliable and valid instruments.

So far we have offered examples of initial entrance into school and of learning difficulty. Let us proceed on our journey to grade eight. This is an important place to pause and consider one very crucial stage in education. It is usually in grade eight that pupils (and their parents) are asked to make a very important decision about a curriculum choice for grade nine. Here again we are concerned with diagnosis and prediction.

We need to know what would be the best curriculum for the pupil to pursue in the immediate future—grades nine through twelve. We need also to have some indication of goals to strive for beyond high school. Good counseling needs to take into account both the immediate and ultimate goals to strive toward. Choice-making which proceeds only on the theory of year-to-year decision making, with next year's choices determined by this year's performance, can be a very wasteful process. We are somewhat handicapped here as yet by an inadequate theory of vocational development—although I hasten to add that much interesting and useful work has been done recently in this field by Roe, Super, and Tiedeman.

Recognizing the lack as yet of a fully acceptable theory of vocational development, it seems to me, as a practitioner, that in order to help a grade eight pupil make his decision we ought to have reliable and valid measures of his achievement, aptitudes, interests, and motivation for learning what we have to offer. Here, too, the problem of comparability enters. It is not enough to have a battery of tests, measuring different traits, standardized on the same population. This is, of course, superior to single tests standardized on different populations. But if two batteries published by different companies and purporting to measure substantially the same traits give different results, what should the practicing counselor think? Is it sufficiently clear that the so-called traits are artificial labels applied to mathematical artifacts? Which battery is really measuring the traits named? Both may have high coefficients of internal consistency.

What do we have in the way of achievement tests for our eighth grade boy which will be helpful to us for purposes of diagnosis and prediction? If we want to estimate the real quality of this boy's ability in writing, we have to fall back largely on teacher judgment. Some bright pupils with high vocabulary, reading, and English usage scores do not seem to be able to express ideas clearly. Is this a matter of emotional blocking? Is it lack of experience? Yet others with similar ability and schooling do write well. Which of our social studies tests really test developed power in historical thinking? What of tests of developed ability in art and music? What of tests of developed mechanical ability?

The development of multi-factor tests of aptitude is a hopeful step towards meeting the need for appraising aptitudes. But much obviously remains to be done as Dr. Super's reviews of existing multi-factor tests in Volume XXXV of the *Personnel and Guidance Journal* clearly indicate. Here the chief concern is chiefly with validity and I recognize that the burden of proof cannot rest wholly with the test authors.

It is unreasonable to expect that tests can be developed in all areas which will have universal validity for all possible situations. The test user must have sufficient sophistication to realize that a given test even with a low validity coefficient may be very valuable for some kinds of use. Also the test user must share the responsibility for determining validity for the particular use to which he wishes to put the test. Perhaps here the test specialist can help the practitioner devise relatively simple methods to help determine validity. Some of this has been done by some publishers.

All of us recognize the inadequacy of interest tests for the eighth grade pupil. Perhaps it is following a will-of-the-wisp to expect that we can ever develop valid interest tests for this age group. I have a hunch that when we have a mature theory of vocational development that we will not find this task as impossible as it seems now. At any rate I present it as a problem to the test specialist.

And the most difficult problem of all has to do with this eighth grade boy's motivation for learning. Can tests be developed which measure motivation for different kinds of learning? Can the "staying power" of a pupil be measured? If a bright boy is not learning well in grade eight, is this a temporary matter? Or is his blocking sufficiently serious so that there is grave doubt as to whether any real desire to do systematic book learning can ever be developed?

We have much advice on all sides today about the responsibility of guidance personnel to encourage intellectually able pupils to go on with further education. Thinking only of the individual pupil, and not of society's needs, how do we know we really ought to do this with a particular pupil? Aside from the very tricky problem of freedom of choice, which is a real concern for counselors, how do we know this bright pupil has or can have developed within him a real desire for the pursuit of learning? Perhaps this is a problem beyond the technology of testing but I present it to you as one of our most perplexing ones.

Examining further the problems in educational and vocational choice-making, it is apparent that more needs to be done to help the counselor as he tries to help a pupil assess his chances for success in a given college or in a particular job. Colleges could be classified in a variety of ways: size, location, nature of student population, and estimate of academic standards. Families of colleges could be established (similar to the effort to establish families of occupations). Prediction with a fair degree of accuracy could then probably be achieved. I understand that C. R. Pace (1) and others have already done some work in this direction. This is a problem which, in particular, confronts the large suburban high school whose graduates go off to school all over the country. It is becoming increasingly a problem of the independent school as well.

There are other stages in the educational process which could be used to illustrate problems in vocational and educational choice-making, but let us examine still another kind of problem which confronts counselors. Here is a 16-year old girl who has just been released from a reform school and placed on parole. She may go to work or return to high school. She chooses to return to school. Should we admit her? How stable is she now? Her offense for commitment had been that of stabbing a boy. Will she be dangerous to our other pupils? Here, too, is an extremely difficult problem.

What of the 16-year old boy who is becoming unmanageable? Is it better to let him leave school? What can testing do to help us with this problem?

II

Up to this point I have presented a few examples highlighting the major problems confronting guidance personnel today. These have to do with matters of personality difficulties, learning difficulties, and educational-vocational choice-making. I am not implying that these are three separate and discrete areas but have given examples which focussed in one area or the other for purposes of illustration.

It is apparent by now that I consider adequate diagnosis and prediction necessary in the work of the counselor and that tests can be useful tools for these purposes. The guidance officer, therefore, needs to know how to select and use tests. The most useful information in helping him select and use tests is that concerned with reliability, validity, and comparability. Let us examine each of these areas from the viewpoint of practicing guidance people. What I have to say now about them is based in part on the results of a questionnaire which I sent to a group of guidance directors and on my own personal experience.*

Reliability and Comparability

All of us would agree that a test is not useful unless it is reliable. When we say this we must also add something about the kind of reliability we mean and its extent, and we must take into account the particular purpose for which we are using the test. Related to this problem of reliability, as tests are used in schools, is the problem of *comparability.*

Reliability and comparability are particularly crucial for tests of mental ability, intelligence, or scholastic aptitude. Try as we will to develop restraint about the value of a score obtained from a group test, too often teachers, administrators, and parents (if they can find out what the result is) treat the score of an intelligence test as an immutable assessment handed down from on high.

To illustrate these problems of reliability and comparability allow me to give you a somewhat extreme example but one that is not rare by any means. The father of a third grade boy in another town came to my office late in the spring to say that he was moving into our city that summer and would our office test his boy for academic aptitude as he felt the boy's ability had been misjudged. One of our psychologists (a trained clinical psychologist) administered a Binet and obtained an I.Q. of 142 which she said she felt was minimal and that if she had really wanted to push the boy he probably would have gone higher. A group test was administered to this boy the next fall and an I.Q. of 102 was obtained. In the sixth grade the next higher level of the same test was administered and an I.Q. of 128 was obtained. In grade 8 a different group test was administered and an I.Q. of 139 resulted. In grade 9 another form of this same test was given and the I.Q. score was 130. Fortunately we give group tests of intelligence at periodic intervals and also occasionally administer individual tests when we feel it necessary. If the individual test and the 8th grade test had never been given we would still have had I.Q.'s of 102 for grade 4, 128 for grade 6, and 130 for grade 9. Obviously there was something completely out of line on the grade 4 score and so a reasonably cautious counselor might assume roughly a probable score of 130. I suspect this boy's true potential is closer to 150. What does this mean in terms of expectation, prediction, and counseling for this boy?

---

*The questionnaire was sent to the members of the Greater Boston Guidance Club (consisting for the most part of guidance directors in East-Central Massachusetts), and a few guidance directors and counselors about the country whom I knew personally. Returns were received from 32 guidance workers in East-Central Massachusetts. Of these 32, 2 were from cities of over 100,000, 14 from cities of 24,000-100,000 and 16 from cities or towns under 25,000. The socio-economic structures of these 32 communities vary considerably as do the backgrounds in experience and training of the respondents. Returns were obtained from 3 large cities, 3 medium-sized cities, and 1 small city elsewhere about the country. In addition, 1 state director of vocational guidance and 2 county directors answered and returned the questionnaires.

All of the tests used in the illustration above have high coefficients of internal consistency. According to the publishers they also correlate highly with one another. One might argue that this was an unstable boy and the differences were caused by instability within the boy. To the best of my knowledge, however, taking into account all we know about him, this boy would be judged as a normal, well-adjusted and stable youngster. What help can the test expert provide counselors for problems of this kind which, I repeat, are not rare?

As guidance practitioners many of us are troubled about the matters of reliability and comparability. If we give a third grade child an I.Q. test we want to be sure that we will get the same I.Q. in grade six, or, if a different one, that the difference is a true difference. We want to be sure that the eighth grade boy will be in the same position on an I.Q. test relative to other pupils when he reaches the twelfth grade. However, follow-up retesting gives wide deviations for too many individuals in spite of publishers' claims to high reliability which are mostly based on the degree of internal consistency. This may actually be more of a problem of comparability rather than reliability since the grade six I.Q. test even though constructed by the same author must of necessity have different content than the grade three one in order to test a higher level of mental ability. Follow-up retesting with alternate forms also gives wide deviations for many individuals, as does retesting with the same instrument several months later.

Would it be possible to provide test-retest correlation coefficients (coefficients of stability)? This is, of course, a difficult technical problem. A true test-retest "r" would be based on giving exactly the same test under exactly the same conditions and to exactly the same pupils. Assuming a negligible amount of learning obtained simply by taking the test, or a large amount of forgetting of the specific items on the test, we cannot avoid change in learning or power caused by the sheer passing of time. I would suggest that the test-retest "r" is more acutely needed with tests which purport to measure power in a given trait or group of traits—more specifically tests which are labelled as mental ability, intelligence, or scholastic aptitude and other so-called aptitude tests. Here the passage of time is not so important, as relative positions of pupils ought not to be seriously affected.

Perhaps this is largely a problem of providing the consumers with more information so that they could choose and use tests more intelligently. Do reliability coefficients vary from one sample to another and, if so, how do the samples differ? Is the standardization group likely to be more heterogeneous than the group which the guidance director is testing? How will this influence reliability? How does teacher administration affect reliability? Were original "r's" obtained with classroom teachers as test administrators? If not, should this be clearly stated and the need for trained administrators stressed if the users wish to obtain reliability comparable to that obtained by the publisher? The increasing tendency to provide instructions for recording results in terms of a range, within which a given score might fluctuate, is all to the good and should be accelerated. This will certainly help in more intelligent use of the test results.

In my judgment this matter of reliability and comparability needs much further attention. Much of the published data on reliability coefficients create too optimistic an impression. And a test publisher obviously has no responsibility for reporting the comparability of his test and that of a rival publisher. The result is that the consumer may unwittingly be the victim not only of the error variance of a particular test, but also of that variance due to the lack of comparability of two tests of the same mental function.

Validity

In the area of validity there seems to be need for clearer and better statements leading to more intelligent use of tests by the consumer. Too often statements on validity are too general and seemingly remote from the day-to-day business of the school or are too much in the nature of a sales pitch. The type of validity (content, concurrent, predictive, or construct) needs to be clarified.

We need to know under what conditions and for what kind of sample the validity data were obtained. For example, is the test of *little* or *no* use with pupils who have reading deficiencies?

More detailed information is needed as to what criteria were used to determine the validity of achievement tests. Which texts or curriculums were used? The guidance director wants to know whether they are the same as or similar to those in his school system, lest we fall into the trap of tests determining curriculum rather than the reverse.

The problem of overlapping levels of achievement test batteries is one that needs examining. This one probably includes reliability, comparability, validity and the good sense of the teachers and counselors. Unfortunately the habit has been created of measuring achievement in terms of grade placement. This habit has been reinforced by some parents (encouraged by some of the hysterical literature attacking public school

104

education) demanding to know exact grade accomplishment of their children. The sixth grade teacher thus eagerly seizes upon an achievement test battery which the publishers tell her can give her equivalent grade placements for her pupils. She properly administers level B which is designed for grades 5 and 6. She is very pleased because many of her children get seventh, eighth, and ninth grade equivalencies and reports this fact happily to the parents of those pupils. The same pupils then enter junior high school where, in October of the seventh grade, the next higher form (level C) of the same achievement battery is administered. What happens? Equivalent grade scores six months to a year less than those obtained in grade 6 are the result. If the school system is lucky it can confine the resulting dispute to the principals of the elementary and junior high schools. If the parents get wind of this difference then their suspicions about soft education in the elementary school are confirmed. Is it the responsibility of the test users not to use scoring methods which are open to all kinds of misinterpretation, or is it that of the specialist to suggest only those methods of scoring which are the least susceptible to misinterpretation?

It would be helpful to have more precise statements of possible use. For example: "This test which we have called *mechanical aptitude* does not by itself predict marks in high school industrial arts courses as well as a test of *verbal aptitude*, but it can be used to improve predictions over those based on verbal aptitude test scores alone." This example illustrates the need for emphasizing caution in accepting titles of tests, and names of "factors" and in the use of profile sheets for making comparisons. Item analysis studies on one group of pupils may be invalid for another due to local phraseology and thus distort test results. For example, in one reading test the phrase "falling fast" occurs and the multiple choice gives rain and storm among other words. Rain is the correct response but in some localities few of us think of rain as falling "fast"—rain falls "hard." A storm might fall fast. As a result although it is one of the earlier and easier questions, in one locality it was missed by 50 per cent of the pupils.

Our previous discussions on educational-vocational choice-making illustrate the need for continuing long-range validity studies and to relate test results more and more closely to such matters as success on the job and in college. As I have indicated before, however, the test user must share some of the burden of determining validity.

Administrative Problems

Allow me to pause a moment here to comment on administrative aspects of testing. By this I mean giving tests, scoring them, recording results, costs, and such matters as format and printing. I am doing this simply because these items have been given a great deal of attention both by practitioners and test specialists. This is not to disparage their importance but merely to make clear that, in my judgment, they do not represent the most important or primary issues which face the guidance director in the use of tests. All test experts are aware of these problems and much attention has been given to them. The less well the tests do what are claimed for them the more important these secondary problems become. I have a hunch that if tests were constructed which really provided answers to fundamental questions, guidance directors, counselors, teachers, and principals might be willing to expend a good deal of time and energy to get those answers. It is only when the tests are of minor value or when there is considerable uncertainty as to what the test scores really do mean, that the tests had better be cheap, easy to administer, score, and record if anyone is to use them. I hasten to add that if we can get tests which really provide answers and are also easy to score, etc., then life would be wonderful. But, speaking very much as an amateur to experts, I have a sneaking notion that there may be an insoluble problem here. It just may not be possible to get the right kinds of answers the easy way. And perhaps we had better stress much more the problem of getting the right kinds of answers and worry a bit less about ease of administration.

III

The preceding discussion on reliability, comparability, and validity stressed the need for more intelligent use of tests. The booklets *Technical Recommendations for Psychological Tests and Diagnostic Techniques* (2) and *Technical Recommendations for Achievement Tests* (3) sponsored in common by A.P.A., A.E.R.A., and N.C.M.U.E. are very helpful for this purpose. Some of the responsibility for bringing this about, however, is directly that of the test specialist in producing better tests, more clearly defined, and with more explicit instructions as to their limitations as well as uses. How can this be brought about?

If there were no need for publishers to concentrate on tests salable to large markets, there would be much benefit in focusing on specific, well-defined populations in specific and well-defined situations. If a test did not have to fit potentially every pupil, even in selected schools, but could be limited in its applicability to pupils with specified characteristics, it would gain considerably in usefulness. At present, too many school people, expecting too much of tests that have been "oversold," may feel disillusioned about the usefulness of testing in general. The test publishers who claim directly or by implication, more than a test in many instances can deliver, are doing themselves a long-range disservice. That the test can deliver for some of the customers does not prevent alienation of others.

Subject matter specialists are consulted to help improve content validity of tests. Why could not committees of guidance people be used to try to help determine the kinds of predictions that are most useful for them? The test publishers could then direct their research to the contribution of test data to such predictions.

For example, we need very badly to have measures of motivation for specific areas of learning. All counselors have had experience with the bright pupil who is not doing well in a particular subject. I recall one very bright boy, in particular, with whom I counseled a few years ago. He was failing in English. It was obvious he had an extensive vocabulary and he spoke in carefully structured sentences such that one could hear the punctuation marks dropping into place. He performed very well on objective, standardized tests in English but was failing largely because he would not hand in any compositions. He protested to me that he was interested in writing, but just somehow couldn't bring himself to sit down and write. His father is a successful free-lance writer and this blocking may have been the result of involved relationships with him. Was there any value in attempting to work with this boy on a supportive level, encouraging him to write? Was this a problem of a profound resistance which could be unraveled, if at all, only by psychotherapy? Or was there simply no interest in writing which was a function of this boy's normal developmental growth and not the by-product of some pathology?

Counselors are concerned also about obtaining reliable and valid tests to measure personality traits such as cooperation, initiative, responsibility. They would like to have a "test of values" suitable for secondary school pupils. They want to know if we can really measure study habits and attitudes.

These are samples of questions and problems which a consulting committee of guidance people might pose to the test specialist to help the test specialist better meet the needs of guidance personnel.

One other approach is for the separate publishing houses to have available field consultants who know measurement and have had considerable practical experience.

Another solution might be to form an independent *Bureau of Measurements Used in Schools* to act in an advisory and selective capacity for the public schools. Such a Bureau could also help in educating parents and the general public as to the function and limitations of tests in an educational program. It could engage in studies of comparability, not for the sake of finding fault, but for providing useful data for the counselor. The Bureau could search for existing problems and needs, perhaps through the device of the Advisory Committee as suggested above, and act as a clearing house for relaying this information to test specialists and publishers.

It is my conviction that we are following a will-of-the-wisp if, as so many practitioners seem to feel desirable, we think we can move towards greater simplicity in this matter of testing. The very uses for which guidance people want tests are complex and becoming more complex. To provide the kinds of tests which will really help a particular guidance person in a particular community do his job well will require an increasingly complex technology. If the local guidance people are to be the determiners of which tests they should use, then the already wide gap which exists between the test specialist and the test user will continue to widen. As the technology increases in complexity or perhaps even changes in some important fundamentals, it will become increasingly difficult for the test user to choose wisely and to use well.

Is the answer to be found in local school systems hiring specialists in measurement? If so, will they determine which tests are to be used for guidance purposes or will they act in a consulting capacity to the guidance director? Can in-service training provide the answer? This is certainly necessary in one form or another if teachers and counselors are to use tests intelligently in their respective roles. If we are to have any in-service training which will have any real impact on the teachers and counselors we had better do something about devising an appropriate training program. Most of our training programs in measurement are devised to produce the expert in measurement We need a consumer's training program that is something else than a watered down version of the program for the prospective expert.

If we really could develop a knowledgeable and sophisticated group of users, the questions posed would tend eventually to be solved. Meanwhile there rests a large ethical responsibility upon the test specialists to make sure that the relatively unsophisticated test users are helped to choose tests wisely and to use them well.

REFERENCES

1. PACE, C. R. "The college environment and its relationship to the prediction of college success." Paper read at the American Psychological Association convention, New York, September 1957.
2. *Technical Recommendations for Psychological Tests and Diagnostic Techniques.* (Prepared by a joint committee of the A.P.A., A.E.R.A., and N.C.M.U.E.) The American Psychological Association, 1333 Sixteenth Street, N.W., Washington 6, D.C.
3. *Technical Recommendations for Achievement Tests.* (Prepared by the Committees on Test Standards of A.E.R.A. and N.C.M.U.E. and endorsed by the A.P.A.) American Educational Research Association, 1201 Sixteenth Street, N.W., Washington 6, D.C.

The Consumer and the Producer

DONALD E. SUPER

A generation ago, A. S. Otis achieved immortality by inventing the self-administering test: you all know, I am sure, that bit of measurement folklore which has it that, anxious to avoid the time-consuming work of individual test administration, he devised the first group test with self-administering directions. Test technology then introduced a second labor-saving device, the self-scoring test, and the names of Clapp and Young became household, or rather, schoolhouse words until machines virtually displaced the "self" with machine scoring. Recognizing the need for local norms, test publishers next made available special forms to assist test consumers in developing their own norms, and the self-norming test came into being. Oddly enough, no test-distributor with merchandising propensities has attempted to capitalize on that term and no test constructor has achieved immortality by associating his name with a so-called self-norming test. And now, Dr. Landy points out, a fourth road to fame, fortune, and the future awaits the test constructor who first devises and markets an instrument which will solve the problem of specific validity, that is, the first self-validating test. Such a test would meet the consumer's need for data on the predictive value of tests for courses in *his* school, for colleges to which *his* students go, for jobs in companies which hire *his* students. Dr. Landy has recognized that publishers cannot accumulate and make these local validity data available to every school which needs them—it is difficult enough to validate for more general types of prediction. The implication: consumers need also to be producers, not of tests, but of test data. Of which more in my closing paragraphs.

The Need for Diagnostic Tests

More regularly than the topic of validity, the need for diagnostic tests comes up time and again in Dr. Landy's talk. But the diagnostic tests which our speaker asks for are not the types of diagnostic tests with which we are familiar, they are something much more complex. We are used to thinking of diagnostic tests of two types: those which assess strengths and weaknesses in a skill subject such as reading or arithmetic, and those which diagnose the mode of personal adjustment. In the former type of test we seek to determine, for example, the arithmetic processes which a pupil has failed to master, so that he may be given the understanding of process or drilled in the use of the process which has given him trouble. In the latter type, we analyze modes of reaction in order to get an understanding of behavior dynamics, so as to use this to guide counseling, psychotherapy, or placement.

But we have heard a request for tests which will diagnose more complex forms of behavior, tests which will tell a counselor whether aptitude, skill, or emotion is the reason for poor performance. The request stems from a better understanding of behavior on the part of guidance workers, from a recognition of the fact that poor achievement may be due to lack of skill in the use of a tool, to lack of aptitude for learning to use the tool, or to emotion which inhibits the student from learning to use or from making use of the tool.

Typically, of course, these complex diagnostic and treatment decisions have been made by the use of a number of different devices which include individual intelligence tests, achievement tests, observations of the pupil at work, personality inventories, projective techniques, and interviews. The use of such methods, as Dr. Landy points out, calls for considerable training on the part of the user and considerable time in which to use them with the student. Current trends in the training of school psychologists, upgrading them from mere givers of individual intelligence tests or processors of group test data to clinicians trained at the doctoral level, reflect the recognition of this fact, even though no comparable upgrading is evident among school counselors. But accompanying this recognition in the case of school psychologists is the awareness of the fact that the number of pupils is increasing more rapidly than the number of school psychologists, that time for more intensive work with more pupils simply is not available. Teachers need treatment suggestions more often and more rapidly than school psychologists can make them.

The cry for a new type of diagnostic test, which will identify the relative roles of aptitude, skill, and emotion, arises from the fact that mass production machines are often developed as substitutes for manpower, to do the missing craftsman's job. Counselors are insufficiently trained in psychology to do a clinician's work and they have workloads which prohibit doing such work even when, as in some notable instances, they acquire the knowledge and skill. But they do have the psychological sophistication to recognize that behavior, even in doing arithmetic or in reading, is complex, and they want tools which will enable them to give teachers the help that teachers want and need.

It probably seems unreasonable to expect a test, or even a test battery, to make a clinical diagnosis. The task seems much too complex for a device or a gadget, appears to call for a human diagnostician. But there was a time when appraising intelligence, or judging mastery of English, or assessing personality, seemed to be the kind of complex clinical task which only human judgment could perform; there was a time when predicting performance seemed also to be something for men, not for machines, to do. But we are all familiar with the fact that tests generally make these evaluations of intelligence, achievement, and personality better than individual judges. We all know that, as Meehl's review showed, they predict success better than judges. Perhaps it is not such a wild idea, after all, to ask for a test, a battery, or a formula which will do a better or a faster job of diagnosing the causes of poor performance than can be done by a psychologist.

The Need for Teams of Psychologists and Technicians

Until someone invents a truly diagnostic test battery and formula, such as Dr. Landy has asked for, we must rely on people. But properly trained people, he has pointed out, are too costly to do the job on the scale on which it must be done. The American Psychological Association acquires about 1,500 new members each year, but only a few of them are fully trained school, clinical, or counseling psychologists, and the recently retooled university training programs will not produce a greatly increased number. School counselors are therefore asked to do the job with the currently available tests. With only one year of training in guidance, and generally not more than two courses in measurement, they have generally had the good sense to recognize their inadequacies as diagnosticians.

What is the answer when the tools of mass production are lacking, when master craftsmen are too few, and journeymen do not have the necessary skills? It consists, of course, of building teams of masters and journeymen, in which the journeymen enable the master craftsman to multiply his services by working with and through them. The qualified psychologist can do less testing, less interviewing, himself. He can rely more on testing done by persons with less extensive and intensive training than his, supervised by him and specializing in the use of certain types of tools. One does not have to be a Ph. D. in order to be able to give a Binet—a technician can do that. The real skill is in the interpretation. A team of psycho-technicians working with one psychologist could do more work, and do it better, than the same number of counselors and psychologists working singly.

It is true that, in the last 15 years, clinical psychology has been unwilling to consider this type of professional structure: it has insisted on the doctorate or nothing. The specialties of school psychology and counseling psychology, on the other hand, have thought in terms of training at both the doctoral and sub-doctoral levels. But they have not applied the team concept, long used and abused by medicine, to their own field. Since abuse of an idea does not necessarily invalidate it, the team concept might well be tried in diagnosis in the educational setting.

Other New Tests

Dr. Landy has asked for some other new tests, particularly tests of values usable at the high school level and tests of motivation. It would be interesting to develop this topic, but time does not permit. Instead, I must move on to more debatable issues.

A Test Consumers' Union

Another need raised by Dr. Landy is that for help in evaluating the large number of available tests and selecting the most appropriate for a particular use, in pooling information about tests, and in educating the public concerning the use of tests. This need is real to me, thanks to a committee on the American Psychological Association which played with the idea of a Bureau of Test Standards while working on a code of ethics, and thereby raised a storm of protest. Work on the ethics of testing led to the establishment of another committee, the function of which was to develop technical standards for psychological tests. Both committees backed away from the idea of any official evaluation of tests; instead, they preferred to clarify standards and stress the development of informed users. But Dr. Landy's suggestion points up another possibility: that of a Test Consumers' Union, comparable to similar organizations which evaluate and report on various types of commodities for their members. Buros' *Yearbook* now does something of this sort, the recent series of articles on multifactor tests in the *Personnel and Guidance Journal* does it in a different way and on a smaller scale, and some test publishers offer consulting services in which public service and merchandising motives are variously mixed. But these are individual or corporative efforts. Perhaps there is room for a test consumers' union made up of school systems organized to pool experience, to share normative and validity data, to protect themselves, and to educate the public. Perhaps we should paraphrase the ancient warning and say, *Caveant mercatores!*

The Training of Test Users

I return, in closing, to the training of test users. Dr. Landy has raised this question very clearly in his statement that most training in testing has been aimed at test specialists. But there are, as I see it, two kinds of test specialists: the psychologists of clinical and counseling persuasions who develop a high degree of diagnostic skill in using tests with individuals, and the psychologists of a metric persuasion who construct the tests. Dr. Landy lumps both types of psychologists together as test specialists in contrasting them with school counselors.

However, it seems to me that what we have heard is actually a plea for the training of more school counselors as test specialists, not as psychometricians, true, but as diagnosticians, in the manner characteristic of clinical and counseling psychologists. But with the shortages which now exist in those fields it does not seem likely that the number of highly qualified test users, trained to work independently, is likely to be greatly increased. I am therefore inclined to re-interpret Dr. Landy's plea as an indication that we need to find better ways of training some psychometrists or psychological examiners at the master's level, able to work as members of psycho-diagnostic teams led by fully qualified school, clinical, or counseling psychologists, and other psychometrists trained as collectors and analyzers of normative and validity data, as members of test development teams led by test constructors.

It is interesting that, in this conference on testing, in this talk on what the guidance director wants from test specialists, is not so much better instruments that are asked for, although we have heard such requests, as better users of instruments which are now available, better ways of making available instruments yield the information which they can be made to contribute. This is not the first time the teaching of testing has come up at these conferences—the 1953 Conference made the subject a major topic on the program. But an examination of the Proceedings confirms Dr. Landy's point: the focus was on testing as measurement, not on testing as understanding people. To many psychologists with a doctorate these may seem to be one and the same thing. But to the user of tests in schools they are not. The question therefore is, could we meet the needs of more people, by teaching testing as the understanding of people rather than as the understanding of measurement?

Appendix

Participants—1957 Invitational Conference on Testing Problems

Abraham, A. A., Florida A & M University
Afflerbach, Janet, Professional Examination Service, New York City
Ahmann, J. Stanley, Cornell University
Albright, Frank S., West Orange (N.J.) Public Schools
Alman, John E., Boston University
Anastasi, Anne, Fordham University
Anderson, Edward L., Educational Testing Service
Anderson, Harry E., Jr., Fort Bliss, Texas
Anderson, Pauline K., New York State Department of Labor
Anderson, Scarvia B., Educational Testing Service
Angoff, William H., Educational Testing Service
Armstrong, Fred G., U. S. Steel Foundation
Aronow, Miriam S., New York City Board of Education
Arsenian, Seth, Springfield (Mass.) College
Atkins, William H., Rutgers University
Baier, Donald E., General Electric Company
Bannon, Charles J., Crosby High School, Waterbury, Conn.
Bardack, Herbert D., New York State Department of Civil Service
Barry, Robert F., Rochester (N.Y.) Board of Education
Bartnik, Robert V., Educational Testing Service
Bates, Margaret A., Educational Testing Service
Beck, Hubert Park, The City College of New York
Bedard, Joseph A., New Britain (Conn.) Public Schools
Bement, Dorothy M., Northampton (Mass.) School for Girls
Bennett, George K., The Psychological Corporation
Bennett, Mrs. George K., The Psychological Corporation
Bennett, Ralph, New York City
Benson, Arthur L., Educational Testing Service
Bentson, T. G., Western Electric Co., Inc.
Berdie, Ralph F., University of Minnesota
Berg, Joel, King Philip School, West Hartford, Conn.
Berger, Bernard, New York City Department of Personnel
Bergesen, B. E., Personnel Press, Inc., Princeton, N. J.
Berne, Ellis, U. S. Department of Health, Education and Welfare
Bicknell, John, New York State Department of Education
Birch, Dorothy L., Educational Testing Service
Bixler, Harold H., Western Carolina College, Cullowhee, N. C.
Bligh, Harold, World Book Company
Blitz, Sidney, New York State Credit Union League
Blommers, Paul, State University of Iowa
Bloom, Benjamin, University of Chicago
Boland, Ruth F., Cambridge (Mass.) Board of Education
Bollenbacher, Joan, Cincinnati (Ohio) Public Schools
Braca, Susan E., Oceanside High School, Forest Hills, New York
Brandt, Hyman, American Occupational Therapy Association, New York City
Bransford, Thomas L., New York State Department of Civil Service
Braun, Louis H., Bronxville (N.Y.) Public Schools
Bretnall, William B., Educational Testing Service
Bridges, Claude F., Washington, D.C.
Brierly, Justin W., Denver (Colo.) Public Schools

Bristow, William H., Bureau of Curriculum Research, New York City
Broderick, J. Lawrence, Y.M.C.A., New York City
Brodinsky, B. P., A. C. Croft Publications
Brolyer, Cecil R., New York State Department of Civil Service
Brown, Fred S., Great Neck (N.Y.) Public Schools
Brown, Leanna, Educational Testing Service
Bryan, Miriam M., Rutgers University
Bryan, J. Ned, National Education Association
Buchheimer, Arnold, New York City Board of Higher Education
Burdock, E. I., Biometrics Research, New York City
Burke, Paul J., Bell Telephone Laboratories.
Burnham, Paul S., Yale University
Buros, Oscar K., Rutgers University
Buros, Mrs. Oscar K., Gryphon Press, Highland Park, N. J.
Butcher, Herbert B., New Jersey Department of Civil Service
Cadwell, Dorothy H. B., Civil Service Commission of Canada
Calvo, Georgia, Educational Testing Service
Capps, Marian P., Virginia State College
Carlson, J. Spencer, University of Oregon
Carlson, Kate, Remedial Education Center, Washington, D. C.
Carroll, John B., Harvard University
Carstater, Eugene D., Bureau of Naval Personnel
Cashman, Jerome P., Archdiocesan Vocational Service, New York City
Cayne, Bernard S., Ginn and Company
Chappell, Bartlett E. S., New York Military Academy
Chauncey, Henry, Educational Testing Service
Choynowski, M., Educational Testing Service
Christensen, Clifford, New York State Department of Education
Churchill, Ruth D., Antioch College
Clark, Pamela, Educational Testing Service
Cleary, Robert, Educational Testing Service
Clemans, William V., National Board of Medical Examiners, Philadelphia
Clendenen, Dorothy, The Psychological Corporation
Cobes, Jeanne M., Educational Testing Service
Coffman, William E., Educational Testing Service
Cohen, Philip S., Montclair (N.J.) State Teachers College
Cole, Joseph W., University of Rochester
Coleman, Elizabeth R., New York City Vocational Education and Extension Board
Collins, Goler L., South Carolina State College
Conrad, Herbert S., U. S. Office of Education
Conway, C. B., Department of Education, Victoria, B. C., Canada
Copeland, Herman A., Pennsylvania State Civil Service Commission
Cornehlsen, J. H., Mahler Associates, New York City
Cory, Charles, Philadelphia Department of Personnel
Crane, Harold L., Jr., Educational Testing Service
Crane, Percy F., University of Maine
Craven, Ethel Case, Polytechnic Institute of Brooklyn
Crawford, J. R., University of Maine
Crosson, Wilhelmina M., Palmer Memorial Institute, Sedalia, North Carolina
Cummings, Mary B., Boston Public Schools
Cureton, Edward E., Knoxville, Tenn.
Cureton, Mrs. Edward E., Knoxville, Tenn.
Curran, Ann M., University of Connecticut
Curtin, Wylma R., Catholic University of America

PARTICIPANTS

Cynamon, Manuel, Brooklyn College
Czukor, John C., New York City Department of Personnel
Dailey, John T., Bureau of Naval Personnel
Daly, Francis J., New York State Department of Education
Damrin, Dora E., Educational Testing Service
Davidson, Helen H., The City College of New York
Dean, E. Douglas, Educational Testing Service
Denison, Violet, Educational Testing Service
Derwin, Edward D., Crosby High School, Waterbury, Connecticut
De Sales, Sister M., Rosary Hill College, Buffalo, New York
Detchen, Lily, Chatham College
de Thomas, Jane, Educational Testing Service
Diamond, Lorraine K., Teachers College, Columbia University
Diamond, M. David, Riverside Hospital, Bronx, New York
Diederich, Paul B., Educational Testing Service
Diggs, Franklin B., New York City Department of Personnel
Dion, Robert, California Test Bureau
Dixon, Robert E., Oberlin College
Dobbin, John E., Educational Testing Service
Doppelt, Jerome E., The Psychological Corporation
Dordick, Mildred, Educational Testing Service
Downes, Margaret C., New York State Department of Civil Service
Dragositz, Anna, Educational Testing Service
Drake, L. E., University of Wisconsin
Dubnick, Lester, Municipal Colleges of New York
Duker, Sam, Brooklyn College
Dunn, Frances E., Brown University
Dyer, Henry S., Educational Testing Service
Ebel, Robert L., Educational Testing Service
Eddy, Robert P., Rutgers University
Edel, Wilbur, New York City Department of Personnel
Edelstein, J. David, New York City Study of Mentally Handicapped Children
Edwards, Winifred, Irvington (N. J.) High School
Eells, Kenneth, University of Illinois
Elder, Samuel J., Mathematics Incorporated, Washington, D. C.
Engelhart, Max D., Chicago City Junior College
Engelhart, Mrs. Max D., Chicago, Ill.
Epstein, Bertram, The City College of New York
Epstein, Marion G., Educational Testing Service
Faulds, Bruce, Educational Testing Service
Fay, Paul J., New York State Department of Civil Service
Feinberg, Mortimer R., The Research Institute of America, Inc., New York City
Feldt, Leonard S., University of Iowa
Fendrick, Paul, Western Electric Company, Inc.
Fenollosa, George M., Houghton Mifflin Company
Fenstermacher, Guy M., Educational Testing Service
Ferguson, John P., The Pingry School, Elizabeth, N. J.
Ferguson, Leonard W., Life Insurance Agency Management Association
Fields, Nina, Croydon Hall Academy, Atlantic Highlands, N. J.
Fifer, Gordon, Test Research Service, Inc., New York City
Findley, Warren G., Atlanta (Ga.) Board of Education
Fink, David R., University of Maine
Finkle, Robert B., Metropolitan Life Insurance Company
Fisher, Evelyn M., New Lincoln School, New York City

Fishman, Joshua A., College Entrance Examination Board

Flanagan, John C., American Institute for Research, Pittsburgh

Fleisch, Sylvia, Boston University

Flesch, Mary H., Educational Testing Service

Follin, Katharine, Remedial Education Center, Washington, D. C.

Forlano, George, Board of Education, Brooklyn, N. Y.

Forlano, Mrs. George, Flushing (N. Y.) Public Schools

Forrester, Gertrude, West Side High School, Newark, N. J.

Frazee, Thelma R., Glen Ridge (N. J.) High School

Freas, Howard J., Jr., Educational Testing Service

Frederiksen, Norman, Educational Testing Service

Freeman, Paul M., Educational Testing Service

French, Benjamin J., New York State Department of Civil Service

French, John W., Educational Testing Service

Fricke, Benno G., University of Michigan

Friedman, Sidney, Bureau of Naval Personnel

Frutchey, Fred P., U. S. Department of Agriculture

Fulton, Renée J., New York City Board of Education

Gallagher, Henrietta, Educational Testing Service

Gardner, Eric F., Syracuse University

Gee, Helen H., Association of American Medical Colleges

Georgia, Sister M., Rosary Hill College, Buffalo, N. Y.

Gerberich, J. Raymond, University of Connecticut

Giddings, Frank, Springfield (Mass.) Trade High School

Glaser, Robert, University of Pittsburgh

Godshalk, Fred I., Educational Testing Service

Goldstein, Leo S., Cornell University Medical College

Goodman, Samuel M., New York State Department of Education

Gordon, Leonard V., U. S. Naval Personnel Research Field Activity, San Diego

Gray, Mrs. Lyle Blaine, Gilman School, Baltimore

Gribbons, Warren, Harvard University

Gribbons, Mrs. Warren, Waltham, Mass.

Grinsted, Alan D., Navy Department

Gross, Reuben H., Jr., A Study of the American High School, New York City

Guerriero, Michael A., The City College of New York

Gulliksen, Harold, Educational Testing Service

Guthrie, George M., Pennsylvania State University

Haagen, C. Hess, Ohio Wesleyan University

Hagen, Elizabeth, Teachers College, Columbia University

Hagin, Rosa A., Irvington (N. J.) Public Schools

Hagman, Elmer R., Greenwich (Conn.) Public Schools

Hall, Alfred, Rutgers University

Hall, Robert G., Manter Hall School, Cambridge, Mass.

Harvey, Philip R., Educational Testing Service

Hastings, J. Thomas, University of Illinois

Hausman, Howard J., National Science Foundation

Hayden, Eric, Public Service, Newark, N. J.

Healy, Ernest A., Jr., Center for Psychological Service, Washington, D. C.

Heaton, Kenneth L., Heaton, Floyd and Watson, Philadelphia

Heil, Louis M., Brooklyn College

Heiser, Ruth B., Glendale, Ohio

Helmick, John S., Educational Testing Service

Hemphill, John K., Educational Testing Service

PARTICIPANTS

Herrick, C. James, Rhode Island College of Education

Heyman, Marshall N., Falls Church, Virginia

Hieronymus, A. N., State University of Iowa

Hitchcock, Arthur A., American Personnel and Guidance Association

Hittinger, William F., State College, Pennsylvania

Hollis, Esther, The Psychological Corporation

Hoopingarner, Newman L., Halesite, L. I., N. Y.

Hopmann, Robert P., Board for Higher Education, Missouri Synod of the Lutheran Church

Horowitz, Milton W., Queens College

Huddleston, Edith M., Educational Testing Service

Hughes, John L., International Business Machines Corporation

Hunt, Thelma, George Washington University

Hunter, Genevieve P., Fordham University

Hypps, Irene C., Washington D. C. Public Schools

Iredell, M. H., National Security Agency

Jackson, Douglas N., Pennsylvania State University

Janeba, Hugo B., Rutherford (N. J.) High School

Jaspen, Nathan, National League for Nursing, Inc., New York City

Johnson, A. Pemberton, Newark (N. J.) College of Engineering

Johnson, Patricia, Educational Testing Service

Juola, Arvo E., Michigan State University

Karack, Goldie Ruth, The City College of New York

Karcher, E. Kenneth, Department of the Army

Karl, Madeline, Brooklyn, N. Y.

Keleher, Gregory C., St. Anselm's College, Manchester, N. H.

Keller, Franklin J., Community Talent Search, National Scholarship Service & Fund for Negro Students, New York City

Kelley, H. Paul, U. S. Naval School of Aviation Medicine, Pensacola, Fla.

Keppich, Charlotte K., Standard Oil Company, (N. J.)

Kern, Donald, University of Bridgeport

Khan, Rafi Z., Federal Public Service Commission, New York City

Kimball, Elisabeth G., Educational Testing Service

King, Jonathan, The Fund for the Advancement of Education

Kingsbury, Mrs. Marion, Remedial Education Center, Washington, D. C.

Kirkpatrick, Forrest H., Bethany (W. Va.) College

Kling, Frederick R., Educational Testing Service

Koch, John C., Jr., Millburn (N. J.) Jr. High School

Komiyama, Eiichi, Tokyo University of Education

Kosmerl, Alice F., Washington, D. C.

Kostick, Max M., State Teachers College at Boston

Krathwohl, David R., Michigan State University

Krauss, Iseli, Macy's New York

Kubis, Joseph F., Fordham University

Kushner, Rose E., Teachers College, Columbia University

Kvaraceus, W. C., Boston University

LaDuke, Charles V., U. S. Armed Forces Institute, Madison, Wis.

Lamke, Tom A., Iowa State Teachers College

Landy, Edward, Newton (Mass.) Public Schools

Lang, Gerhart, The City College of New York

Langmuir, C. R., The Psychological Corporation

Lannholm, G. V., Educational Testing Service

Layton, Wilbur L., University of Minnesota

LeBold, William K., Purdue University

118

Lebow, Daniel B., New York City Department of Personnel
Lee, Louise F., Remedial Education Center, Washington, D. C.
Lennon, Roger T., World Book Company
Levine, Richard, Educational Testing Service
Lindquist, E. F., State University of Iowa
Litterick, W. S., The Harley School, Rochester, N. Y.
Litty, Frank, New Jersey State Department of Civil Service
Lohman, Maurice A., New York State Department of Education
Loiselle, H. George, Dade County Board of Public Instruction, Miami, Fla.
Long, Louis, The City College of New York
Lord, Frederick, Educational Testing Service
Lord, Shirley H., Educational Testing Service
Loree, M. Ray, Louisiana State University
Lorge, Irving, Teachers College, Columbia University
Lorge, Sarah W., James Monroe High School, New York City
Losi, Carrie R., Board of Education, Newark, N. J.
Loughran, George A., St. Francis College, Brooklyn, N. Y.
Lusk, Louis T., Norwalk, Conn.
Lutz, Orpha M., Montclair (N. J.) State Teachers College
Lyman, Howard B., University of Cincinnati
Lyons, William A., New York State Department of Education
MacKay, James L., South San Antonio (Texas) Schools
Mahony, Thomas L., New Jersey State Department of Civil Service
Manuel, Herschel T., University of Texas
Marston, Helen M., Educational Testing Service
Masia, Bertram B., Science Research Associates
Mathews, Chester O., Ohio Wesleyan University
Maxwell, Sheena, Edinburgh, Scotland
Maxwell, James, Teachers College, Columbia University
McCann, Forbes E., Philadelphia Department of Personnel
McCord, Richard B., Philadelphia Department of Personnel
McCracken, Charles W., Trenton (N.J.) State Teachers College
McIntire, Paul H., University of New Hampshire
McKenzie, Francis W., Darien (Conn.) Board of Education
McLaughlin, Kenneth F., Florida State University
McManus, Leo F., Jr., University of Connecticut
McQuitty, John V., University of Florida
Meade, Martin J., Fordham University
Medley, Donald M., Municipal Colleges of New York City
Melville, S. D., Educational Testing Service
Merry, Robert W., Harvard University
Merwin, Jack C., Syracuse University
Messick, Samuel, Educational Testing Service
Metz, Elliott, Little Neck, New York
Micheli, Gene, Metropolitan Life Insurance Company
Miles, Matthew B., Horace Mann-Lincoln Institute, New York City
Milholland, John E., University of Michigan
Millard, Kenneth A., Department of Defense
Miller, Bernard S., A Study of the American High School, New York City
Miller, Harold T., Houghton Mifflin Company
Miller, William N., U. S. Government
Mitchell, Mrs. Blythe C., World Book Company

119

PARTICIPANTS

Mitchell, Col., Department of the Army

Mitchell, Robert H., Newton (Mass.) High School

Mittman, Arthur, State University of Iowa

Mitzel, Harold E., The City College of New York

Mollenkopf, William G., Procter & Gamble

Moore, James W., California State Scholarship Commission

Morgan, Henry H., The Psychological Corporation

Morrison, J. Cayce, The Puerto Rican Study

Mosely, Russell, Wisconsin State Department of Education

Muirhead, David B., Michigan State University

Muirhead, Peter P., New York State Department of Education

Munger, A. M., Standard Oil Co. (N.J.)

Myers, Sheldon S., Educational Testing Service

Nally, Thomas, University of Rhode Island

Nedelsky, Leo, University of Chicago

Nevin, Margaret, Educational Testing Service

Noll, Victor H., Michigan State University

North, Robert [illegible], Educational Records Bureau

Noslow, Samuel, Reading Institute of Boston

Nosow, Sigmund, Michigan State University

Olsen, Marjorie, Educational Testing Service

Orellana, Egidio, Instituto Pedagogico, Santiago, Chile

Orleans, Joseph B., George Washington High School, New York City

Oscarson, Donald, The Taft School, Watertown, Conn.

Oxtoby, Toby, Technical Reports, Inc., New York City

Packard, Albert G., Department of Education, Baltimore, Md.

Palmer, Orville, Educational Testing Service

Palmer, Osmond E., Michigan State University

Partington, Dorothy B., Educational Testing Service

Pashalian, Siroon, Community Service Society, New York City

Patton, James B., Jr., Virginia State Department of Education

Peabody, Elizabeth R., Groton School, Groton, Mass.

Perlman, Mildred, New York City Department of Personnel

Perloff, Robert, Science Research Associates

Perry, William D., University of North Carolina

Peters, Frank R., Ohio State University

Peterson, Carla A., Educational Testing Service

Peterson, Donald A., Life Insurance Agency Management Association

Philpott, Emily L., Queens College

Pierson, George A., Queens College

Pitcher, Barbara, Educational Testing Service

Pollack, Norman C., New York State Department of Civil Service

Preston, Braxton, Educational Testing Service

Price, Leah, Uniondale High School, Valley Stream, N. Y.

Quinn, John S., Jr., World Book Company

Rabineau. Louis, Pratt Institute, Brooklyn

Rabinowitz, William, New York City Board of Higher Education

Randall, Harvey, New York State Department of Civil Service

Rankin, Paul T., Detroit Public Schools

Ransohoff, Priscilla B., Consultant Associates, Inc., Long Branch, N. J.

Rapparlie, John H., Owens-Illinois Glass Company

Raskin, Evelyn, Brooklyn College

Ray, William S., Pennsylvania State University

Read, Thomas, Maumee Valley (Ohio) Country Day School
Reed, Anna K., New York State Department of Civil Service
Reiner, Harry, New York City Department of Personnel
Remmers, H. H., Purdue University
Reuter, William H., Educational Testing Service
Rich, Jeanne, Rutherford (N. J.) High School
Richards, Roger E., The Psychological Corporation
Ricks, James H., Jr., The Psychological Corporation
Riedl, Norbert F., Educational Testing Service
Rimalover, Jack K., Educational Testing Service
Robbins, Irving, Queens College
Robertson, Marguerite E., University of Connecticut
Roblyer, William A., The Choate School, Wallingford, Conn.
Rosinski, Edwin F., University of Buffalo
Ross, Mrs. C. A., The Pingry School, Elizabeth, N. J.
Rossi, O. D., Schenectady (N. Y.) Public Schools
Rulon, P. J., Harvard University
Rundquist, E. A., Alexandria, Va.
Runkel, Philip J., University of Illinois
Sait, Edward, Rensselaer Polytechnic Institute
Salzberg, Florence, Queens College
Samler, Joseph, Veterans Administration, Washington, D. C.
Sanders, Edward, Pomona College, Claremont, Calif.
Sanford, Nevitt, Vassar College
Saunders, David R., Educational Testing Service
Sawin, Enoch I., U. S. Air Force ROTC Headquarters, Montgomery, Ala.
Scates, Alice Y., U. S. Office of Education
Schapiro, Harold B., Young & Rubicam, Inc.
Schepers, J. M., Princeton University
Schrader, William B., Educational Testing Service
Schulman, Frances H., Institute for Career Guidance, New York City
Schulman, Jay, Institute for Career Guidance, New York City
Schultz, Douglas, Pennsylvania State University
Schwartz, Alfred, University of Delaware
Schwartz, Joe, Mitchel Air Force Base
Scott, C. Winfield, Rutgers University
Scott, Winifred Starbuck, New Brunswick, N. J.
Seashore, Harold G., The Psychological Corporation
Seibel, Dean W., Educational Testing Service
Setzer, Charles J., New York City Department of Personnel
Sforza, Richard F., New York State Department of Civil Service
Shanner, William M., California Test Bureau
Shapan, Norma, Board of Child Welfare, Trenton, N. J.
Sharp, Catherine G., Educational Testing Service
Shaycoft, Marion F., American Institute for Research, Washington, D. C.
Shea, Lester T., Hauppauge School System, L. I., N. Y.
Shear, Bruce, New York State Department of Education
Shields, Mary, National League for Nursing, Inc., New York City
Shimberg, Benjamin, Educational Testing Service
Simon, Dorothy M., New York City Department of Personnel
Simpson, Elizabeth A., Illinois Institute of Technology
Sitgreaves, Rosedith, Teachers College, Columbia University
Smit, Jo Anne, Arlington, Virginia
Smith, Alexander F., New Haven (Conn.) State Teachers College
Smith, Allan B., University of Connecticut

PARTICIPANTS

Smith, Ann Z., Educational Testing Service

Smith, Denzel D., Office of Naval Research

Smith, Robert E., Educational Testing Service

Smith, William Reed, University of Utah

Snodgrass, Robert, Educational Testing Service

Sohmer, Kenneth D., Millburn (N. J.) High School

Solomon, Robert, Educational Testing Service

Souther, Mary Taylor, Tower Hill School, Wilmington, Del.

Spaney, Emma, Queens College

Spaulding, Geraldine, Durham, North Carolina

Speer, George S., Illinois Institute of Technology

Spencer, Richard E., Jackson (Mich.) Public Schools

Spencer, W. Douglas, New York City

Spencer, Wesley G., Cambridge, Mass.

Stahl, Suzanne, Educational Testing Service

Stake, Robert, Educational Testing Service

Stallman, Frederick B., New York Telephone Company

Statler, Charles R., University of Chicago

Stecklein, John E., University of Minnesota

Stein, Morris I., University of Chicago

Stephan, Frederick F., Princeton University

Stewart, Mary, Institute of Physical Medicine and Rehabilitation, New York University

Stice, Glen, Educational Testing Service

Stodola, Quentin C., Educational Testing Service

Stokes, Thomas M., Metropolitan Life Insurance Company

Stone, Paul T., Huntingdon College, Montgomery, Ala.

Stoughton, Robert W., Connecticut State Department of Education

Strong, Edward K., Jr., Stanford University

Stuart, William, Educational Testing Service

Stulbaum, Harold, Elmhurst, New York

Super, Donald E., Teachers College, Columbia University

Svendsen, Johan, Teachers College, Norway

Swanson, Edward O., University of Minnesota

Swift, Everett L., The Peddie School, Hightstown, N. J.

Swineford, Frances, Educational Testing Service

Syferd, Janet, Educational Testing Service

Tanofsky, Robert, Metropolitan Life Insurance Company

Tansey, Gertrude, Lakewood (N. J.) Senior High School

Taylor, Justine, Educational Testing Service

Taylor, Samuel J., Philadelphia Department of Personnel

Terkel, Meyer, Yeshiva University, New York City

Terral, J. E., Educational Testing Service

Thomas, William F., University of Wisconsin

Thorndike, Robert L., Teachers College, Columbia University

Tinkle, Jack W., Mitchel Air Force Base

Torre, Mottram, Columbia University

Trail, Stanley M., Rhode Island College of Education

Traxler, Arthur E., Educational Records Bureau

Triggs, Frances, Committee on Diagnostic Reading Tests, Inc., New York City

Tucker, Ledyard R, Educational Testing Service

Turnbull, William W., Educational Testing Service

Tyler, Leona E., University of Oregon

Urban, Hugh B., Pennsylvania State University

URQUHART, Helen L., Brown University
VALENTINE, John A., Educational Testing Service
VALLEY, John R., Educational Testing Service
VICKERY, K. N., Clemson College, Clemson, S. C.
VOSS, Harold A., Office of Naval Research, Port Washington, N. Y.
WADELL, Blandena C., World Book Company
WAGNER, E. Paul, State Teachers College, Bloomsburg, Pa.
WAHLGREN, Hardy L., State Teachers College, Geneseo, N. Y.
WALDMAN, John, Pace College, New York City
WALLACE, Wimburn L., The Psychological Corporation
WALLER, Raymond L., Allentown (Pa.) Public Schools
WALLMARK, Madeline, Educational Testing Service
WALSH, John J., Boston College
WALTON, Wesley W., Educational Testing Service
WARD, Annie W., University of Tennessee
WATKINS, Betty J., Educational Testing Service
WATKINS, Richard W., Educational Testing Service
WATSON, Walter S., The Cooper Union
WEBSTER, Harold, Vassar College
WEISS, Eleanor S., Educational Testing Service
WEISS, Joseph, Polytechnic Institute of Brooklyn
WEITZ, Henry, Duke University
WEITZMAN, Ronald, Educational Testing Service
WELLCK, A. A., University of New Mexico
WESMAN, Alexander G., The Psychological Corporation
WEST, Elmer D., American Council on Education
WHIGHAM, E. L., Wilmington (Del.) Board of Public Education
WHITNEY, Alfred G., Life Insurance Agency Management Association
WIENER, Solomon, New York City Department of Personnel
WILKE, Marguerite M., Greenwich (Conn.) Board of Education
WILKE, Walter H., New York University
WILKINS, Walter L., St. Louis University
WILKS, S. S., Princeton University
WILLARD, Richard W., Massachusetts Institute of Technology
WILLEMIN, Louis P., Department of the Army
WILLEY, Clarence F., Norwich University, Northfield, Vt.
WILLIAMS, Roger K., Morgan State College, Baltimore, Md.
WILSON, Kenneth M., Princeton University
WILSON, Robert C., Reed College, Portland, Ore.
WINANS, S. David, New Jersey State Department of Education
WINGO, Alfred L., Virginia State Department of Education
WINIEWICZ, C. S., U. S. Naval Examining Center, Great Lakes, Ill.
WINTERBOTTOM, John A., Educational Testing Service
WIRSIG, Jane, Educational Testing Service
WIRSIG, Woodrow, Educational Testing Service
WOMER, Frank B., University of Michigan
WOOD, Ray G., Ohio State Department of Education
WOODBURY, Max A., New York University
WOOLLATT, Lorne H., Baltimore (Md.) Public Schools
WRIGHT, Wilbur H., State Teachers College, Geneseo, N. Y.
WRIGHTSTONE, J. Wayne, New York City Board of Education
YATES, Vivian M., New York City
YOUNG, Donna S., University of South Carolina
ZALKIND, Sheldon S., New York University

PARTICIPANTS

Zeidner, Joseph, Department of the Army

Zeigler, Martin L., Pennsylvania State University

Zubin, Joseph, New York State Psychiatric Institute, New York City